Representing Transcription Factor Dimers By Using Forked-Position Weight Matrices

by

©Aida Ghayour Khiavi

A Dissertation submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of

M.Sc.

Department of Computer Science

Memorial University of Newfoundland

May 2021

St. John's Newfoundland

Abstract

Position Weight Matrices (PWMs) and sequence logos are one of the most popular tools among researchers for modelling and visualizing Transcription Factor (TF) Binding Sites (TFBS). The PWM based models predict a single DNA sequence as a reference TFBS for a specific TF, based on experimentally determined sequence information. One of the standard assays for characterizing the TFBS of one TF on a genomic-wide scale is called ChIP-Seq. The Chromatin Immunoprecipitation (ChIP) method uses TF-specific antibodies to capture protein:DNA complexes, followed by high-throughput sequencing of the bound DNA sequences (Seq). These experiments are applied in a controlled manner to target only one TF at each run, thus describing TFBSs of a single TF of interest. This approach is proven to be imprecise because many TFs (e.g. Leucine Zippers) tend to bind to the DNA as homodimers or heterodimers. Hence, the ChIP-seq assay will obtain the entire set of dimer complexes of a target TF (homodimers and heterodimers); and merge the captured information into a single PWM which subsequently will lead to an imprecise description of the TFBS. The TFBS constructed by the mixture of homodimers and heterodimers will result in a model with two halves: a conserved part (binding sites of the TF of interest) and a degenerated part (representing a mixture of the binding sites of TF's partners). Current PWMs (or Sequence Logos) seem inadequate to represent TF dimer binding sites since they fail to represent the TF's binding dynamic and disregard the alteration in

sequence preference caused by different dimer partners of the given TF. To tackle this problem, we introduce an R library named Forked Position Weight Matrix (FPWM), which provides the user with variant functionalities to generate a more precise PWM that adapts to TF dimers by forking it into the co-factors of the main TF. The FPWM enhances TFBS prediction's power and allows the biologists to have a more precise interpretation of cell context by providing a more expressive model of TFBSs. The FPWM is less susceptible to false-positives and is a more precise way to represent dimer TFBSs, which introduces a new standard in dimer and TFBSs analysis.

General Summary

If we consider the DNA a book written using 4 alphabets of A, T, C and G, then each gene is a sentence. The final expression each sentence makes depends on factors such as the context of the book or the usage of meaningless but necessary elements like punctuations. It is proven now that 98% of human DNA does not encode for a specific function but has the role of regulating genes' expression. Some experimental and computational methods try to see which set of genes interplay with which of these regulatory factors. One of the methods for this purpose is ChIP-Seq, which is designed to draw this relation between only one factor and its associated genes. In this thesis, we explain how ChIP-Seq can miss information in some cases since some factors work as two partners or dimers. Thus we proposed a model that takes this dimerization into account, thus representing more precise data about the dimer factors and their association with DNA.

Acknowledgements

I would like to thank the following people who have helped me undertake this research:

My supervisors, Drs. Touati Benoukraf and Hamid Usefi for their support;

The Departments of Computer Science, Human Genetics and Mathematics for input and financial support throughout this MSc. program;

Members of our lab, Roberto Tirado-Magallanes and Lin Xiao Xuan for their assistance with integration of this work within initial in house projects;

Dr. Ann Dorward, for her support and helpful guidance and Dr. Sevtap Savas, her faith in graduate students.

Zoha Rabei, for her kindness, strength and wisdom;

And eventually, my mom, dad and brother for their love, Sam, Ellie, Hasti and Sahar for their existence, and Simanto for his heart.

Statement of Co-Authorship

A significant amount of job has been done before this project, which allowed me to develop current work with the help of those resources. Several people were involved from different backgrounds prior to the beginning of this project, which should be acknowledged as follow:

My special thanks go to **Lin Xiao Xuan** from Cancer Science Institute of Singapore (CIS), a former member of our lab and developer of the MethMotif Database and TFregulomeR compendium. The WGBS and ChIP-Seq detests, were prepossessed and available in the MethMotif Database, and we were able to immediately export and integrate them, with the help of TFregulomeR() API.

I also would like to thank a valuable member of our lab, **Roberto Tirado-Magallanes** from CSI who was a great contributor to the design of secondary analytical steps and optimization of the scripts for big data processing (which was required for submission of paper in high impact factor journals.)

Eventually, I would like to recognize the contributors who supported me in developing this thesis and allowed me to integrate this toolkit into the new version of the MethMotif database (MethMotif 2021):

Sudhakar Jha (from CSI) , **Morgane Thomas-Chollier and Prof. Denis Thieffry** from Institut de Biologie de l'École Normale Supérieure (IBENS), for their useful observations as the R-SAT suit developer team, which was employed in this project

to evaluate our final results and making analytical comparisons.

Matthew Dyre, a valuable member of our lab for his great work in providing the FPWM toolkit in MethMotif's website for public access.

Dr. **Touati Benoukraf**, my supervisor and the leader of Benoukraf-lab, for his strategic plans throughout all these projects, and Dr. **Hamid Usefi**, my co-supervisor from department of Mathematics for his help with the thesis.

Publications

This project have been published in several formats so far, as follows:

1. The manuscript was initially submitted to the journal of Bioinformatics (impact factor of 5.61), and currently, we are working on the revisions. The list of co-authors and their contributions is reported in the previous section as a statement of co-authorship.

2. This work has been represented as a subdivision of MethMotif in two conferences, namely "Applied Bioinformatics in Life Sciences (3rd edition)", held in Leuven, Belgium, on February 13-14, 2020.

Figure 1: Our work has been selected for a poster presentation during the VIB Conference Applied Bioinformatics in Life Sciences (3rd edition).

3. The current project was also represented at 28th conference on "Intelligent Systems for Molecular Biology (ISMB)," on July 16-13, 2020. The conference was located in Montreal and held virtually due to known circumstances.

Figure 2: The FPWM was presented at the visual conference on Intelligent Systems for Molecular Biology (ISMB), 28th.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

FPWM	Forked Position Weight Matrix
PWM	Position Weight Matrix
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
DNA	Deoxyribonucleic Acid
Seq	Sequencing
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
RNA	Ribonucleic Acid
A	Adenine
T	Thymine
G	Guanine
C	Cytosine
U	Uracil
ENCODE	Encyclopedia of DNA Elements
RT-PCR	reverse Transcription(RT) Polymerase Chain Reaction
RIP-Chip	RNA immunoprecipitation chip
DNase	Deoxyribonuclease

FAIRE	Formaldehyde-Assisted Isolation of Regulatory Element
RRBS	Reduced Representation Bisulfite
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag
NGS	Next-Generation Sequencing
PCR	Polymerase Chain Reaction
MACS	Model-based Analysis of ChIP-Seq
SICER	Spatial Clustering for Identification of ChIP-Enriched Regions
Q	Quality
ID	Identity Document
ASCII	American Standard Code for Information Interchange
SAM	Sequence Alignment Map
BAM	Binary Alignment Map
PFM	Position Frequency Matrix
PCM	Position Count Matrix
HOCOMOCO	Homo Sapiens COmprehensive MOdel COllection
RSAT	Regulatory Sequence Analysis Tool
CpG	C-Phosphate-G
WGBS	Whole Genome bisulfite Sequencing
GTRD	Gene Transcription Regulation Database
bZIP	Basic leucine ZIpper
BED	Browser Extensible Data
TRANSFAC	TRANScrption FACtor (data base)
ABLS	Applied Bioinformatics in Life Science
ISMB	IntelligentSystems for Molecular Biology

Chapter 1

Research Background

In this document, the interdisciplinary field of bioinformatics has been investigated. Regarding the research area and the importance of understanding this report's purpose, a basic introduction to molecular biology and common terms and phrases is provided first. Then the research background and related projects will be described in this chapter.

1.1 DNA, Genes and Chromosomes

Based on an article published by "Watson and Crick" in 1953[1], DNA (DeoxyriboNucleic Acid) is a double-helix polymer that base units, called nucleotides, from each strand bind to each other in a complementary manner. This results in a ladder-shaped molecule with base pairs of Adenine-Thymine or Guanine-Cytosine[2]. A gene, on the other hand, can be defined as "a unit of DNA that is usually located on a chromosome controlling the development of one or more traits and is the basic unit by which genetic information is passed from parent to offspring[3]." Genes are the functional units of DNA, traditionally considered as protein-coding genes but also recognized as functional RNAs (RiboNucleic Acids).

DNA is a very long molecule that must be compressed into small packages to fit into the cells within organism's body. The free form of DNA and protein combination is called chromatin. In order to compress this, DNA coils around the protein structures (Histones) many times, forming a compact structure like beads, called nucleosome (see 1.1).

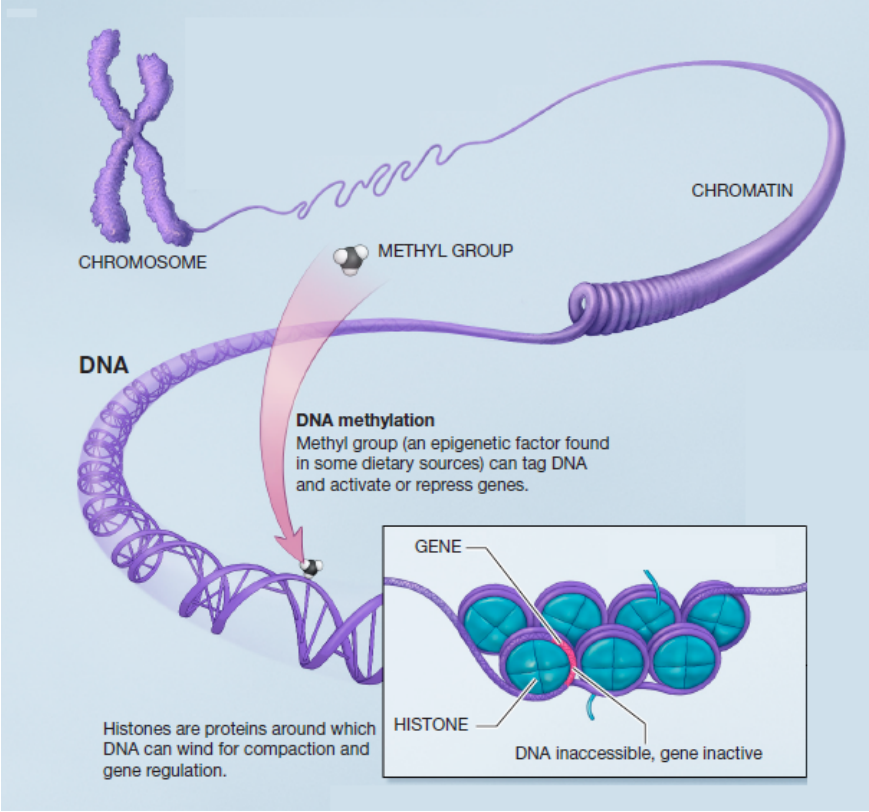


Figure 1.1: Chromosomal DNA is compacted by coiling around histons forming nucleosomes. This structure is compressed further to make the chromatin of a chromosome. As can be seen, this spacial structure, makes DNA exposed to some environmental elements (such as methyl structures) or keeps them inaccessible. *This image is a modified version of a pictures was adopted from U.S. National Institute of Health[4] on October 2020. (permission of access granted since image is available for public use with citation.)*

Chromatins then condense further, constructing a characteristic structure called a chromosome. The structure of the chromosome and how it is constructed, as described, is depicted in Figure 1.1.

Human beings have 46 chromosomes, 23 coming from each parent. A part of genetic researches focuses on understanding the role of DNA in orchestrating the development of cells, tissues and organs with the help of many available genome sequence data. The level of access to DNA sequences, is a determinant of how genes are expressed and that is controlled through a dynamic protein-DNA structural organization (e.g. chromatin). Once the accessibility is achieved, genes can be exposed to their environment for inter-playing with with other structures such as transcription factors.

1.2 Gene Expression

Cellular DNA holds the sequence information to specify the gene expression pattern of each cell. The means of transmitting the information encoded in the DNA is described as the central dogma of biology. The central dogma refers to the direction of the information flow of a mechanism in which genetic information within the DNA results in a functional product such as protein[5]. RNA is a molecule highly similar to DNA except that it has only one strand instead of two (making its structure relatively unstable) with a Uracil (U) instead of Thiamin (T). RNA makes a complimentary copy of the DNA sequence in order to transmit the information encoded within DNA[6]. This information transition (DNA -> RNA -> Protein) has two levels: transcription and translation. Transcription is the crucial step determining which genes will be expressed (or switched on) while in translation, the RNA made in the previous level is translated to extract the information encoding the resulting protein. The product of gene expression leads to specific functions, which perform various tasks, and specific regions within the DNA, have regulatory effect on this expression. Almost 10% of the whole genome is constructed from genes, while 90% of it is non-gene regions that control genes regulation. So, in order for these genes to be recognized and expressed,

a set of regulatory proteins identify gene regions by binding to specific sites on DNA, then initiate transcription to start the gene expression[7].

1.3 Transcription Factors and their Binding Sites

Transcription Factors (TFs) are proteins that attach to Binding Sites (TFBSs) of their preference on DNA and regulate the gene expression by interacting with the regulatory element of that very gene. So, regarding their determining role in the gene expression mechanism, it is necessary to find their binding sites on the genome to predict the way it is going to affect transcription[8]. A single TF can bind to many locations on DNA. The TFBSs can simultaneously target multiple TFs, leading to a complex regulatory effect at binding sites. Although these sites may be distant from the affected gene (up to 2 million base pairs [9]), they can associate with the genes through chromatin loops. Some TFs can direct remodelling of chromatin and reposition the nucleosomes, which subsequently opens the door for even more TF recruitment. TFs also can prevent this from happening by keeping nucleosomes from relocating[10]. TFs are categorized into multiple classes, regarding their structure and how they bind to the DNA sequence. According to Figure 1.2, currently, there are ten superfamilies of TFs. The three largest ones are called: Basic domains, Helix-Turn-Helix domains and Zinc-coordinating DNA-binding domains. In Figure 1.3, examples of these three superfamilies in interaction with a DNA sequence is depicted.

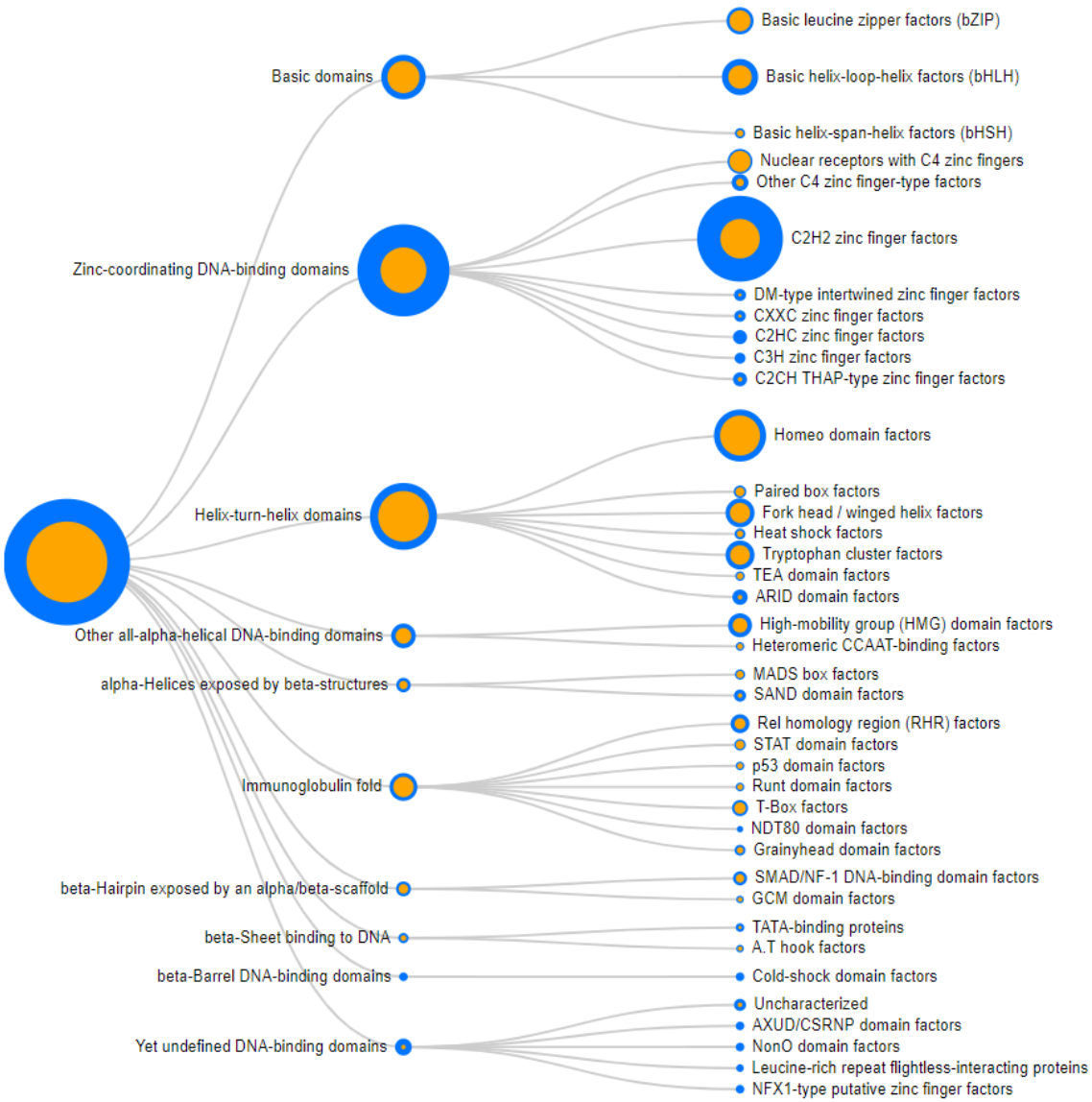


Figure 1.2: Current set of TFs are categorized into ten super families, each of which containing multiple families. *The image is captured from HOCOMOCO's website [11] without any alterations. The permission is granted from Oxford University Press on October 2020*

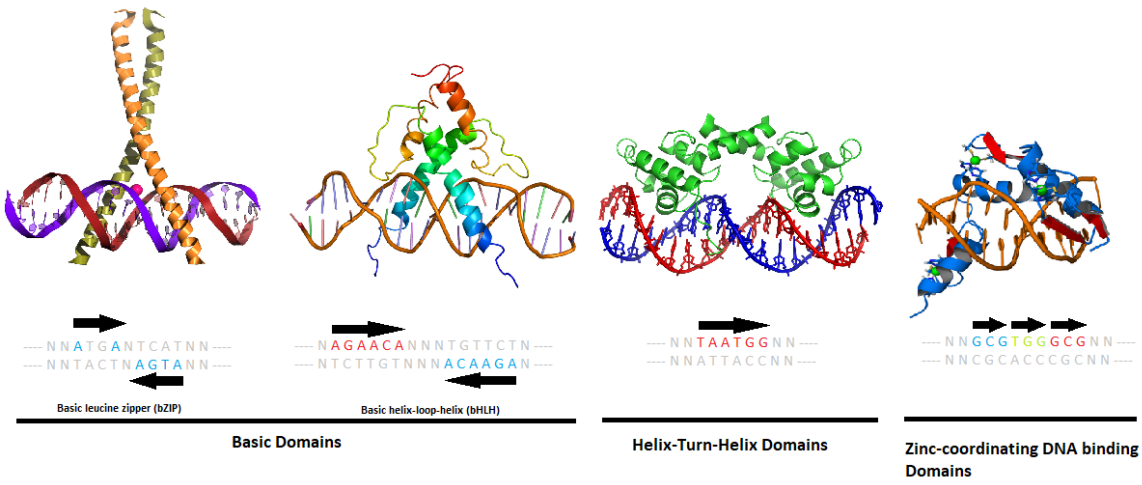


Figure 1.3: Representation of 3 TF super families in interaction with DNA sequence. *This picture is adopted then modified (from [12]) with direct permission from author.*

1.4 Mapping Transcription Factor Binding Sites

After the Human Genome Project's successful execution with the main goal of describing the entire sequence of base pairs that make up human DNA, the time came to identify its functioning regions. After attempting to find and allocate each region to one functional product, it was confirmed that only 1% of 3.3 billion nucleotides code for proteins. Mammals have retained non-coding regions within their DNA sequence and that many of gene-related disease or traits are associated with non-coding regions [13]. Therefore, the importance of identifying coding-regions and drawing a relation between them and traits or diseases seemed to be more important than ever. By the advancement of technology and development of more precise methods of genome wise study, the interest in analyzing these non-coding regions in DNA and their potential role is rising rapidly[14]. Given that 99% of the human genome appears to be non-coding, the question is what these regions are responsible for? Recently a project was initiated in which researchers try to further define the role of these regions in

the entire genome's functionality, and that although the non-coding regions do not directly code for protein, they have a significant role in regulating those regions that do code for proteins.

1.4.1 ENCODE Consortium

Understanding the significance of non-protein coding regions of the human genome was the motivation behind a broad consortium project named Encyclopedia of DNA Elements (ENCODE). ENCODE project was started in 2003 by the National Human Genome Research Institute. They gathered professionals skilled in computation to approach the problem of "Identifying active regions of the DNA" with a combination of high-throughput methods[15]. The ENCODE project stresses on quality of generated data. All the data generation stages and its associated information should follow a set of defined standards to be easily and systematically used and regenerated. For this purpose, high-throughput technologies have been a great help, as they have made the overwhelming task of data management remarkably straightforward. The central Data Coordination Center is exclusively assigned to revising, collecting, and spreading the data sets after a quality check. The resulting data collection and the analysis based on those (such as graphs and data files) are publicly accessible through the website. Overall, ENCODE follows a well-defined set of protocols to generate an integrated set of data that is easy to handle and study. Following this, different types of data are generated by the consortium, which is scientifically useful. A number of these data sets are as follow[16]:

1. Genes and Transcripts: The project's main goal is to annotate different DNA sites to the list of the transcriptional products in hand. Accurate annotation of all the non-coding sites or pseudogenes has not been an easy task. It is worth mentioning that multiple algorithms have been developed that undertake this

task and do it fast and automatically. However, the manual annotation done by a human is still the most reliable approach. The ENCODE consortium's priority is to rely mostly on a manual approach with the help of automatic modelling of the genes and transcriptional products that can be revised by traditional approaches.

2. Cis-Regulatory Regions: The regions close to the gene that regulate the gene expression are called Cis-regulatory elements. These regions may include a couple of elements, namely promoters, enhancers, silencers, etc. One of the principal works of the ENCODE project includes locating Cis-elements and identifying the TFs that bind to them to help scientists quickly investigate the behaviour of regulatory elements and their impact on the expression of a particular gene.

3. Additional Data Types: The ENCODE project takes other data types under consideration to complement some other ongoing projects or produce benchmark data for public use. For example, they assemble data sets of DNA methylation. Methylation is one of the processes that a DNA may undergo. To put it simply, it is discovered that the addition of a methylation group to the DNA sequence is an inheritable epigenetic phenomenon. Regarding this fact, ENCODE tries to calculate the level of methylation at each DNA sequence of interest. This will be described in further details in the following sections.

The ENCODE consortium employs various biochemical assays shown in Figure 1.4 to generate data and study DNA sequence. Other scientists also use most of these experimental approaches in order to collect data. A number of these assays are as follows[17]:

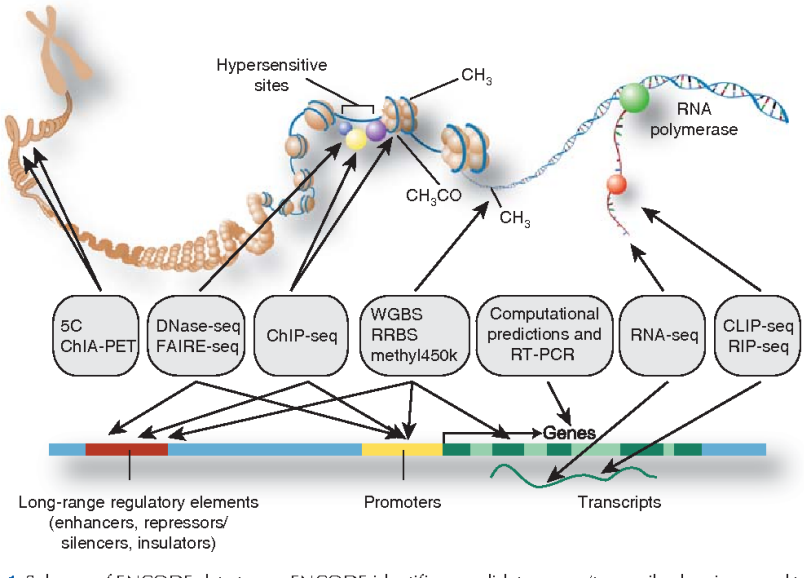


Figure 1.4: ENCODE: Encyclopedia of DNA Elements: employed techniques and methods scheme.
From [17], with the permission of use from Cold Spring Harbor Laboratory Press Bookstore, granted on October 2020.

1. RT-PCR: Reverse Transcription(RT) Polymerase Chain Reaction is a technique used to analyze gene regulation patterns by measuring the amount of RNA of interest[18].

2. RNA-seq: This approach consists of making a library of Complementary-DNA (cDNA) from a set of RNAs followed by high-throughput sequencing, which subsequently will form a map of the transcriptional arrangements and possibly each gene's expression rate[19].

3. RIP-Chip: In RNA immunoprecipitation chip method, the RNA of interest, along with the proteins bound to it, is isolated through an assay, so it helps the researchers to identify a set of RNAs that may have similar protein performance or resembling roles[20]. In some experimental cases, this approach has led to identifying motifs of the 6-8 nucleotide length and was observed to be remarkably

specific and low on error[21].

4. DNase-seq: An enzyme named Deoxyribonuclease I (DNase I) exists within our cells that digest DNA chaotically. However, due to chromatin's characteristic structure, this enzyme cannot contact all the DNA sequence regions. DNase I-hypersensitive site sequencing (DNase-seq) is a method to locate those regions where digestion occurs, which interprets it as locating the regulatory elements[22].

5. FAIRE-seq: Formaldehyde-Assisted Isolation of Regulatory Elements, allows researchers to isolate regulatory elements in the whole genome in an efficient way. This assay, followed by a high-throughput sequencing method, results in a highly efficient method for identifying regulatory regions on DNA sequence[23].

6. RRBS: Reduced Representation Bisulfite Sequencing is a method to study the DNA methylation level on a genome. This model has developed as an alternative for WGBS and employs a relatively reduced region of DNA sequence to represent the whole methylation pattern[24].

7. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET): Using this method, scientists can investigate genome-wide interactions of the chromatin and how regulatory elements interact with one another through different stages of cell differentiation or development. CHIA-Pet helps draw the relation between regulatory proteins and their binding site throughout the genome[25].

Chip-Seq and Whole Genome bisulfite Sequencing (WGBS) will be explained in further detail as they are the core of our work.

1.4.2 Whole Genome bisulfite Sequencing

Methylation of DNA refers to a chemical interaction in which methyl groups are added to the fifth position of the Cytosine structure (see Figure 1.5).

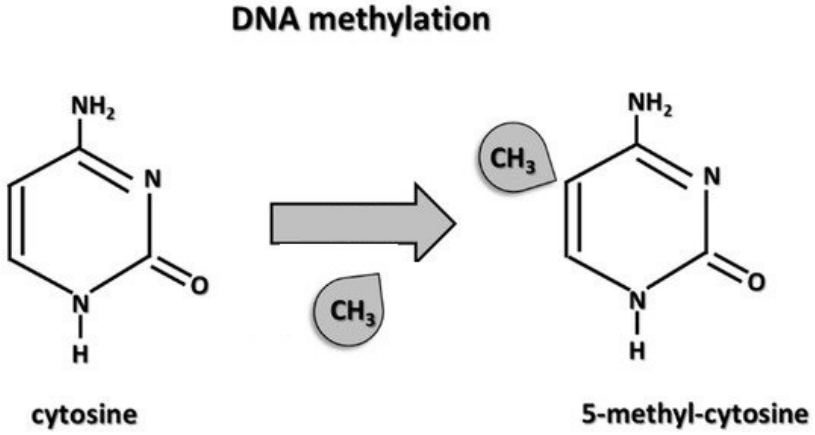


Figure 1.5: Attachment of a structure of methyl group to fifth position of cytosine nucleotide, results in methylation.
This image is a modified version from [26], with a permission under Creative Commons Attribution (CC BY) license available at:
`http://creativecommons.org/licenses/by/4.0/` .

This phenomenon is stable and has a significant impact on gene expression throughout the cell's developmental stages, and is considered one of the most influential factors in the dynamic of TF's binding preference[27]. It has been reported that methylation can inhibit TFs from binding to the regulatory elements; however, recent studies show that this is not a general rule as some TFs bind only to methylated DNA, and some show no specific alteration in behaviour[27]. Since the research foundation of this thesis focuses on the dynamic nature of a given TF's binding site and hence its functionality in different contexts; it is vital to understand how methylation is measured and represented. To measure the level of methylation, the DNA is treated with a chemical named bisulphite that converts all the unmethylated Cytosine nucleotides to uracil while the rest remains unchanged. Coupling this assay with high throughput

sequencing facilitates the entire procedure. By doing this, it is possible to observe the level of methylation at each Cytosine, which gives the researchers a broad insight into the gene's behaviour[28]. In one of our in house projects (MethMotif[27]), the level of methylation profile of each TFBS is intuitively represented as a bar chart, representing the state of methylation on the given TFBS. The level of methylation at each position is categorized into three groups based on methylation score percentages: first <10% (i.e. homogenously hypomethylated); secondly, methylation scores >90% (i.e. homogenously hypermethylated) and finally, methylation scores ranging from 10% to 90% (i.e. heterogeneously methylated)[27]. An example of a methylation profile for a given TFBS coupled with its DNA profile, studied in a cell-specific context, is represented in Figure 1.6.

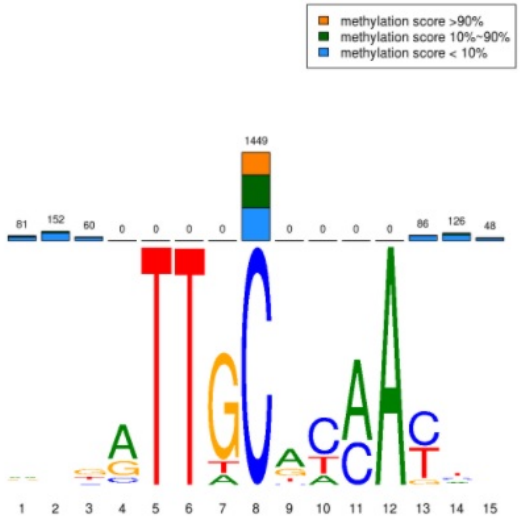


Figure 1.6: An example of Methylation Profile couple with sequence motif profile. The three intervals expressing the level of methylation is shown in different colors. *This Figure is adapted from [27]*, on October 2020 with the permission of the authors.

1.4.3 ChIP and ChIP-seq

Chromatin Immunoprecipitation

ChIP is a highly efficient technique to study the association of a target protein (or TF) with a DNA sequence. [29]. ChIP consists of multiple stages (see Figure 1.7). Each of these stages should follow district instructions in order to achieve the highest accuracy and efficiency at the end. Generally, ChIP starts with fixing the protein structure bound to a DNA sequence at their bound position. This is done by applying formaldehyde onto the sample, which forms a covalent link between protein structure and the DNA sequence. "Formaldehyde is a reversible protein-DNA cross-linking agent that serves to fix or “preserve” the protein-DNA interactions occurring in the cell [30]." Once the protein structures are fixed in their place, the DNA string and those complexes bound to it undergo a mechanical or chemical shearing procedure, which breaks the whole segment into short fragments while the bound sequences remain protected. The next step is immunoprecipitation.The target protein:DNA complex, is selectively isolated by using a protein-specific antibody. This helps the enrichment of a selection of fragments that are attached to protein structures. It is worth mentioning that the selective isolation of a particular protein: DNA fragment helps ChIP procedure detect continuous and distant fragments of the DNA targeted by the protein of interest. Note that this strategy has its drawbacks, such as high reliance on the specificity of the applied antibody or the amount of protein of interest that exists in the sample also being a snapshot in time. It is strongly recommended to have a highly specific antibody (to decrease the experimental noise) along with a large amount of expressed protein of interest (to increase the accuracy of analysis). After the immunoprecipitation phase, the DNA fragments and the protein attached to them go through a reversed cross-linking stage in which their attached protein is

washed away. In the end, what left is purified DNA fragments that were targeted (around 150 bp) by the TF of interest. It is only by analyzing this pool of selected DNA fragments that we can monitor the behaviour of TF of interest and see how it affects the gene regulation[31].

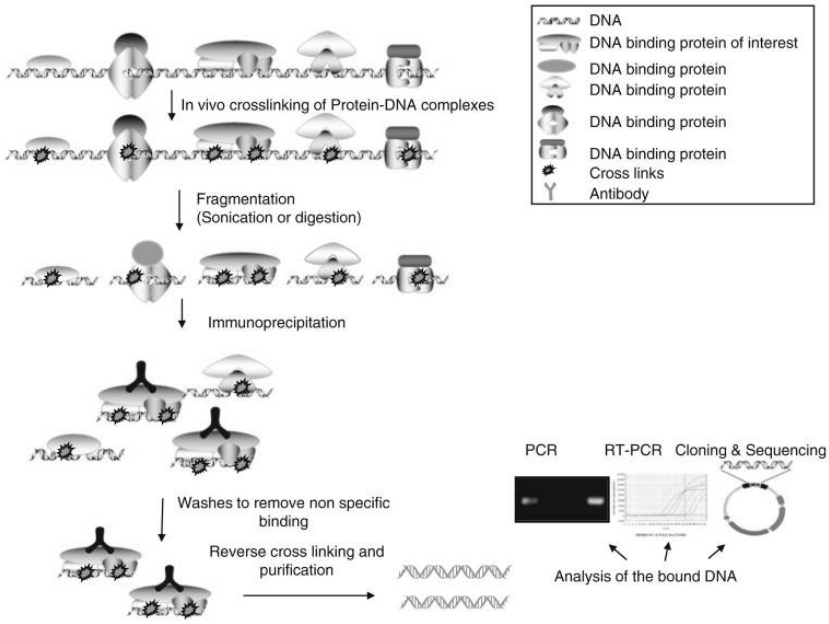


Figure 1.7: Major stages of a typical ChIP assay for library construction.
This Figure is adapted from [31] under permission of use from the Springer Nature publisher.

ChIP Followed by Sequencing (ChIP-Seq)

The term DNA sequencing refers to the procedure of finding the order of nucleotides within the genome. The latest technologies that enable researchers to sequence millions of DNA fragments in only one set of experiments allowed researchers to perform extensive and highly time-consuming projects. Next-Generation Sequencing (NGS) or high-throughput sequencing is one of the novel ideas of recent years which refers to set of sequencing technologies that allows researchers to sequence DNA or RNA faster and cheaper than traditional methods. NGSs have been employed in many

projects and experiments such as RNA sequencing, characterization of the DNase I sensitive sites on the genome, or finding a new type of small RNAs [32]. Regarding these remarkable achievements with NGS's help, it is expected that with the development of the third generation of sequencing technologies, significant discoveries and experiments will be executed even more. ChIP assay, followed by sequencing, was one of the most remarkable utilization of NGS. In ChIP-Seq, the sequencing is limited to the DNA sequences of interest only. The coupling of ChIP assay with NGS results in a highly efficient and accurate study (if the determinants were chosen adequately), which also benefits from the observation's vast range. Therefore, the generated data by this method is highly reliable. However, the ChIP-seq method's properties, such as each read's short length (around 25–32 base pairs), may not be quite preferable for some other applications, but it is entirely compatible with ChIP-Seq technology. ChIP-Seq presents a more accurate characterization of TFs and enhancer regions and measures more numbers of these elements per each experiment's run. The high accuracy and reliability of ChIP-seq, along with its ability to study a larger portion of the genome, makes it a very suitable and recommended approach for studying protein-DNA interactions. The general workflow of the ChIP-seq may include some main stages (as shown in Figure 1.8), namely quality control, alignment, peak calling and motif predication in following order[32]:

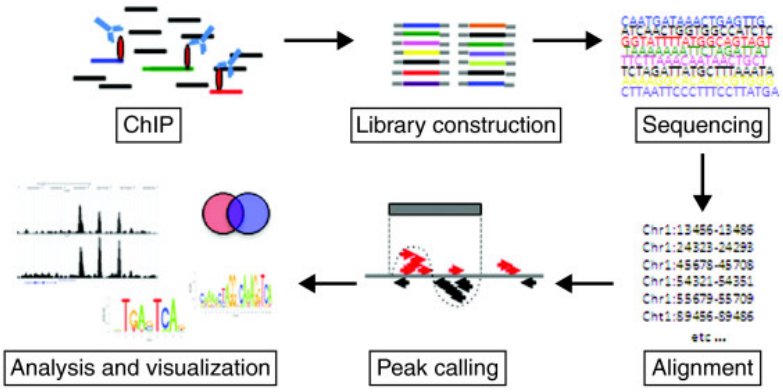


Figure 1.8: The workflow of Data generation with the ChIP-seq experiment and the procedure which ends to Motif Prediction.
The figured is adapted from [33], on October 2020 with the permission of use and distribute with citation.

1. Library Construction: After Chromatin Immunoprecipitation, the extracted DNA fragments undergo a couple of revising and modifications to construct a library that is ready to be sequenced. For this purpose, adapter sequences are adjusted to the fragments to prepare them for the next step. It is worth mentioning that the main challenge in library construction is the amount of DNA. As mentioned before, since the number of sequences is limited to the sequences of interest, there might be a phase in which the number of fragments is increased by applying the Polymerase Chain Reaction (PCR) method. After the construction of a proper library, fragments are ready to be aligned[34].

2. Alignment: In the field of genetics, alignment is the procedure of finding the location of a collection of sequences by relating them to their source reference in order to find their location of origin on the reference genome (refer to Figure 1.9). The result of the alignment stage is the generation of a data source that can later be studied for related genomic features, aside from locating, such as DNA sequence comparisons and gene expression rate[35].

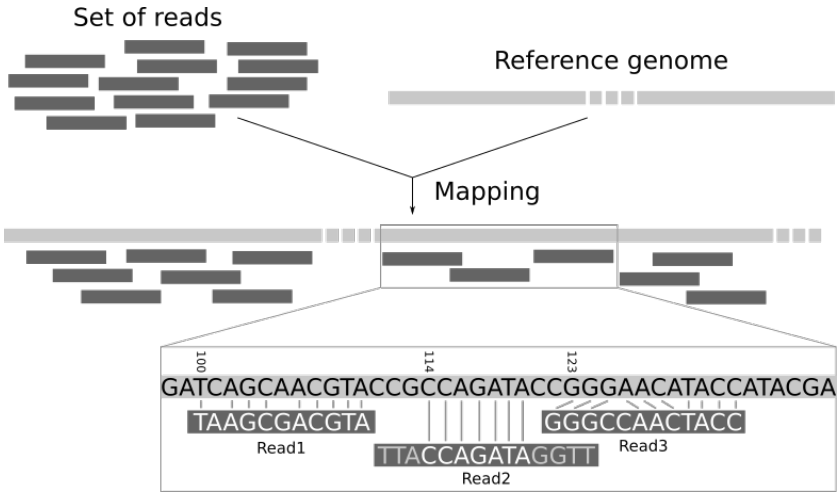


Figure 1.9: Set of reads are aligned with the help of the reference gennome resulting the map. First read is aligned at position 100, second one at 114 and the third read at position 123. Note than insertion, deletions and mismatching at some positions are possible scenarios.
This Figure is from [36], accessed on October 2020. The content of this website is open for adapt and distribution under the Creative Commons Attribution 4.0 International License.

3. Peak Calling: After the alignment of DNA fragments, it is necessary to start the computational approach to discover the TF of interest's target fragment. For this purpose, those regions of the genome from which the most DNA fragments are originated should be validly detected. "Peak calling" studies the stack of reads formed after aligning many fragments to the reference genome. These stacks or "Peaks" may correspond to a TF of interest (TFBS). The peak calling procedure includes two significant stages: Detecting the location that many of the reads are piled upon each other, followed by validating that the detected stack is indeed a signal from a TF's binding Site, and not a false signal[37]. However, the levels of works are more than just these two. The peak calling algorithm highly relies on the mapping layer's performance. So it is necessary to make sure that these two tools are compatible. The peak callers normally go through several layers, namely: read shifting, background estimation, identifi-

cation of enriched peaks, significance analysis and removal of artifacts. First, the aligned reads (normally 150-300 bps) are shifted to merge both strands' data, thus identifying the sub-sequence high probably involved in protein: DNA interaction. The size of shifting can be set based on the size of the fragments used in library construction, which can be done either experimentally or by approximating from sequence data. Note that the comparison of these two can also be a way to work on the quality since the reads' ratio coming from opposite strands is expected to be almost 1. The next step would be identifying peaks. This is done either by setting a threshold and monitoring the value coming from each peak or detecting locations with minimum enrichment compared to the background observed within a sliding window. After this phase, the significance of peaks is determined. Most of the toolkits deliver a P-value for each peak; many of them, on the other hand, rely on the height of the peak and/or compare the enrichment to the background to rank the peaks. Finally, the artifact data should be removed. First, a set of peaks that contain few reads are removed with this assumption that they are the result of PCR amplification. Then those peaks that have a high imbalance between reads coming from opposite strands are deleted. After this phase, a list of peaks is delivered to the user as the set of validated peaks. The choice of peak caller highly relies on the type of experiment that is being run. For example, some of the peak callers work better for the transcription factors. There are several popular peak callers, namely Model-based Analysis of ChIP-Seq (MACS)[38], which is a commonly used tool, especially for TFs. It is relatively easy to work with due to little required adjustments. However, sometimes, the user is not interested in the peaks but the type of data that can range kilo or megabases of the genome. Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) [39] is developed for those

types of studies, for example, chromatin modification.

Quality determinants

Several experimental determinants can impact the quality of ChIP-Seq assay results, as follow: The first one is the bias that may arise in the genomic coverage of reads, potentially manipulating the core signal. The next one is the library type, which can be constructed using paired-end or single-end sequencing. It is reported that paired-end libraries are more suitable for the proposes, such as identifying alternative splice isoforms or discovering chimeric transcripts [40]. The third determinant to name is the sequencing depth of ChIP and input chromatin samples (constructed by fragmenting or enzymatic digestion of chromatin extracts), engaged as a control for background signal. And finally, the type of strategy that is used for peak calling[40]. It is stated that in a certain level of standards, the analytical algorithms and the choice of computational tools have a more significant impact on the final result than experimental factors[41]. For example, it is reported that the minimum number of cells that are required to operate a ChIP-Seq experiment can be reduced by fine-tuning the experimental parameters, such as the quality of the antibody. The quality of the final result can be improved by optimizing the concentration of formaldehyde used for cross-linking the proteins and DNA regarding the type of cell and antibody in use. Also, high-quality outcomes can be obtained by precise control of DNA shearing and fragmentation of correct size[42]. The fragmentation size should be short enough to result in narrower peaks initially, but not too short that they are not mapped accurately. Regarding the abovementioned, the importance of computational analysis of ChIP-Seq assay, which is the goal of this project evident. In the following, the Bioinformatic strategies around ChIP-Seq will be described.

1.5 Bioinformatics Analysis of ChIP-Seq data Sets

ChIP-seq data sets are a valuable data source only if they are properly analyzed using the appropriate analytical tools. The analysis of all data starts with managing the retrieved data. The first step to comprehensively manage similar data sets is to transform them into the same structure to facilitate data exchange and software compatibility. As was described earlier, data collected from a ChIP-Seq experiment represents sequencing information retrieved by the second stage of the assay (high-throughput sequencing). However, the abundance of file formats that are generally not well-defined or flexible enough has been one of the biggest bioinformatics challenges. Many formats are do not respect the standard protocol that everyone should be following, yet are highly employed because of their strength. In the case of DNA sequencing, one of the commonly used data formats, which is highly utilized for data exchange among variant tools, is FASTQ. The FASTQ data format is an upgraded version of a data format with the same properties, but with an extra column for the quality score to provide some information about the reliability of each nucleotide recorded. For example, in the Phred method, some lookup tables are constructed by recording the correctly validated sequences from sequence traces[43]. These tables are hard-coded and are used by Phred to look up and calculate some factors based on each peak's resolution and shape[43]. Phred scores are originated as an algorithmic strategy, in which some parameters such as peak resolution and shape are connected to a known sequence precision stored as a lookup table. These lookup tables are generated by analyzing the factors related to specific sequencing chemistry of a large experimental data set of known accuracy. Q scores are generated based on a logarithmic approach to base calling error probabilities (P) such as:

$$Q = -10log_{10}P. \tag{1.1}$$

For example, a Phred quality score of 20, implies that in 100 sequences base units, there is one error. So, in other words, the probability of an incorrect base call is 1 in 100, which means base call accuracy is 99%[44]. It is necessary to keep this qualities in a systematic format. FASTQ is a text file, including four mandatory lines. The first line starts with an @, followed by an ID. Although the ID is not arbitrary, optional information can follow it. Then comes raw letters of the sequence, immediately following each other without a tab or space. For instance, in DNA, consequent A, T, C, and Gs follow each other to represent the sequenced fragment. In the third line, a "+" symbol comes to signal the end of the sequence (it can also be followed by optional information). In the final row, each letter's quality score is presented in the second row. For this purpose, ASCII characters are assigned to each level of quality, and those characters are used to represent the quality of each letter of the second row's sequence. As expected, the number of characters in the fourth line is equal to the number of letters in the second line[45], as depicted in figure1.11. Note that "the byte representing quality starts from 0x21 (lowest quality; '!' in ASCII) to 0x7e (highest quality; ' ' in ASCII). Here are the quality value characters in left-to-right increasing order of quality (ASCII)"[46]:

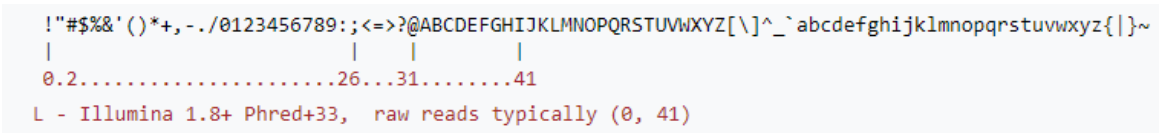


Figure 1.10: ASCII characters corresponding to different quality values in FASTQ format.In earlier protocols, the " " was the last character. However, in last protocol of Illumina, quality range needs only 42 first character, reflecting score from 0 to 41. *This text is adapted from [46]*, and is permitted to use and distribute under Creative Commons Attribution-ShareAlike 3.0 Unported License.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)

Figure 1.11: A typical FASTQ data file contracted from three sections: Record ID, Sequence and quality recorded for each letter of sequence, using ASCII characters. *This Figure is adapted from [45], no permission required.*

The FASTQ data format is used to represent a sequence and how reliable each letter has been recorded as described in the workflow of ChIP-Seq after sequencing comes alignment. The aligned set of sequences need to be represented in a data format for further analysis. One of the current alignment data formats compatible with all kinds of sequences and aligners is Sequence Alignment Map (SAM). This data format consists of two parts: Header and Alignment. Basically, SAM keeps a record of sequences, along with the data obtained from aligning them against a reference genome. All the lines in alignment sections start with an @ and contain obligatory 11 fields that can be increased by optional ones. There 11 columns are QNAME (Read Name), FLAG (SAM flag which can be decoded), RNAME (contig name or * for unmapped), POS (mapped position of base 1 of a read on the reference sequence), MAPQ (mapping quality), CIGAR (CIGAR string describing insertions and deletions), RNEXT (Name of mate), PNEXT (Position of mate), TLEN (Template length), SEQ (Read Sequence), QUAL (Read Quality) and TAGS, which can hold optional information in TAG:TYPE:VALUE format [47].

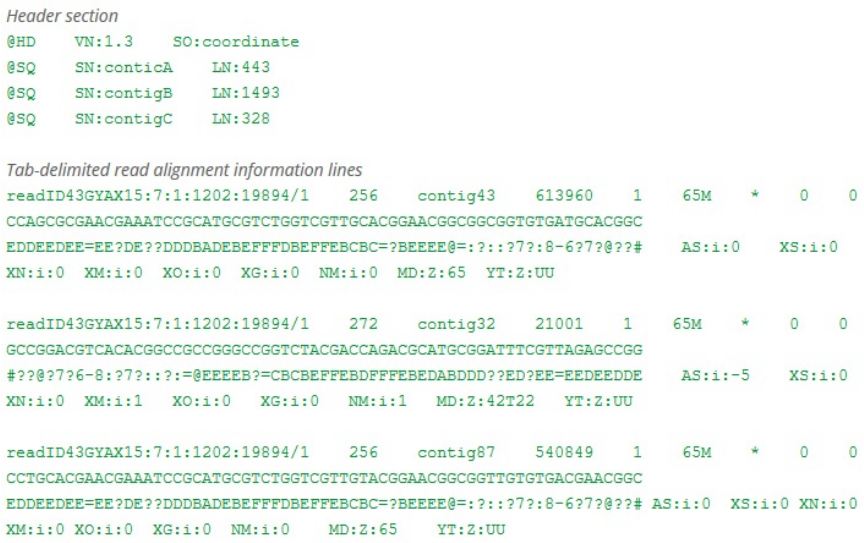


Figure 1.12: A typical SAM data file containing 11 mandatory columns and number of optional elements under TAGS.
This is a modified version adapted from [47], with open access (No permission required).

SAM can be converted to a binary form for sufficiency proposes, named Binary Alignment/Map (BAM). BAM contains the same information as its associated SAM, except that it represents it in binary format[48]. With achieving a comprehensive set of sequences aligned against the reference genome, we can move to the next step, detecting and verifying those locations on the genome that most of the reads originate from, using peak calling algorithms. These are the regions that the protein of interest may have been interacting with the DNA sequence. There are many methods and approaches for peak detection, and each has its own definitions and regulations. However, before taking a look at the general concept of the peak, it is worth understanding the retrieved data better. The nature of the ChIP-seq experiment each tag can only represent one of the DNA strands) results in the dependency of tags on the strand of the DNA. As explained before, the DNA molecule is constructed from two strands that complement each other. To distinguish between these two, each end of the DNA molecule is named (5' and 3'). So, mapping each of the sequence fragments

onto the reference genome is done concerning the reference sequence and DNA fragment's direction[49]. This strand-dependent mapping of the tags will form a diagram of the form, as depicted in Figure 1.13.

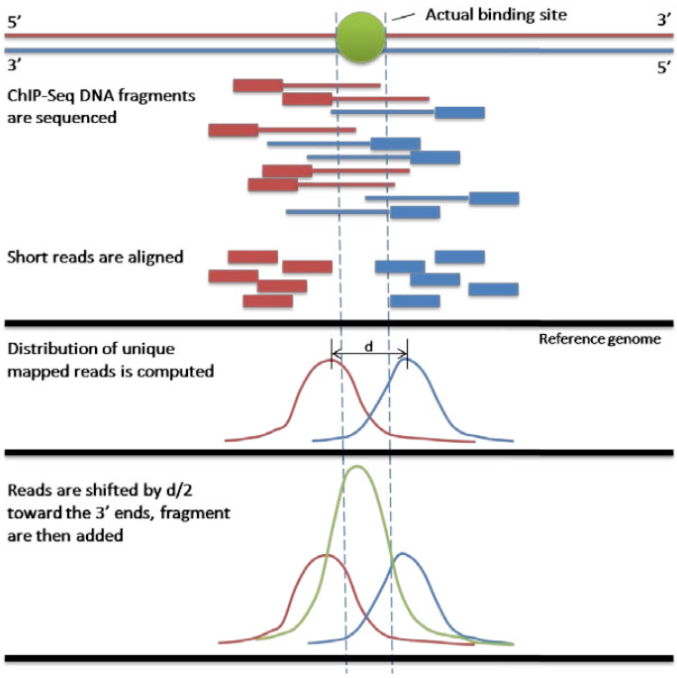


Figure 1.13: Reads are sequenced from both end in ChIP-seq procedure. Therefore, a TFBS signals two set of peaks each at one of the directions. If the reads are shifted towards each other half of the length of d, then one single peak will be formed indicating potential TFBS's region.
This Figure is adapted from [50] (openly accessible) on October 2020, without any alteration.

After mapping, the tags' distribution diagram is rendered, forming two peaks, each responding to one of the DNA strands. The gap between the extremum point of two peaks corresponds to the length of the produced tags; based on this, all the tags are shifted to the 3' end of the DNA strand they are mapped on, which results in a merge of two first peaks. The new peak of the distribution diagram indicates the actual TFBS[49]. After locating the TFBS, the region will be focused on. As it was depicted, the peak locates a region upon which many tags have been plied. However,

the concentration of tags at all the positions within the TFBS is not the same, which owes to the fact that TF binds to the DNA in a sequence-specific manner. It means that TFs can bind to many sequences that are not the same, but they seem to follow a pattern.

Sequences sharing similar pattern are found within the DNA, which are called Motifs . Motifs are believed to have a specific biological role, such as indicating Binding Site for TFs. Many motifs have been discovered all through the years of study regarding their essential role in gene expression. The sequence-specific behaviour of TFs results in different degrees of binding to a motif, which directly impacts the gene expression level. For example, in 1975, it was discovered that the "TATAAT box" is a motif located before the Transcription Initiation Site, which is highly conserved in Escherichia coli promoters. This motif, along with another motif, forms the RNA Polymerase II binding site, which initiates transcription. Although the conservation level at each position is high, finding a promoter with the same sequence is hard. Many of the promoters match 7 -9 out of the total 12 base-pairs, determining the level of the gene's activity[51]. In order to describe these motifs (TFBSs in our study) and to analyze how strongly they appear in variant conditions, it is necessary to drive a model from the data gained from ChIP-Seq (or any other technology). It is only by such a model that one can have an accurate understanding of a motif's general structure.

1.5.1 TFBS Representation, Scanning and Prediction

TFBS are short fragments of DNA that are located in regulatory locations. The precise description of these Binding Sites is the goal of many computational or experimental projects in genetics. There are many experimental and computational approaches for this purposed number, respectively, in [52] and [53]. Although so

much effort is put into the description of TFBSs, which has discovered many of those, the number of identified TFBSs is not even close to the proposed real number. For example, the ENCODE project has described almost 200 TFs in less than 100 human cell lines. Regarding the study's massive scale, the sensitiveness of the case of study (the nucleotide level resolution of each defined TFBS) employment of efficient methods to predict TBFBs is demanded. Computational approaches have been applied extensively and resulted in favourable outcomes. These methods can vary from simple pattern matching strategies to highly complicated ones. Our focus in this study is on pattern matching methods based on "PWMs" because of their popularity and efficiency. These models are used to predict a TFBS by presenting a candidate sequence using a model that is constructed from the data of experimentally discovered TFBSs. These models are widely employed because of their accuracy and simplicity despite them being around for decades[54].

Generally speaking, Position Weight Matrices are scanned against the DNA to recognize a TFBS. Variant computer-based tools are developed in order to recognize a TFBS by scanning PWM against DNA. Which either predict a unique TFBS or classify them into a cluster. Come of the approaches for PWM scanning are reviewed in [55] and [56]. After discovering a regulatory sequence motif, all the target genes regulated by that motif, which may bind to a TF, should be recognized. Consensus sequences represent the Binding Site's characteristics, but they do not reflect accurate information about nucleotide alternation at each of TFBS's positions. So, for degenerate sequence preferences of TF's, PWM is employed[56].

Position Weight Matrix

Position Weight Matrix (PWM) is the central tool in most of the motif prediction applications nowadays. In the mildest case, its input includes a collection of aligned

sequences (as shown in Figure 1.14) and declaration of background frequencies. This method's output will be the PWM and statistical data and information content for the site and motif[57]. Position Weight Matrix is a 4-row matrix (in case of representing Nucleotide sequences) with the number of columns equal to the length of the site that it represents. To construct a Position Weight Matrix, a simple Position Frequency Matrix (PFM) is built as represented in 1.15. The PFM depicts the total number of a specific nucleotide (A, T, C, and G) that has been repeated at each position in a set of aligned sequences. However, for computational efficiency, this matrix needs to be normalized. For that purpose, a table of probabilities is constructed by dividing the elements of PFM by the number of aligned sequences, which results in the sum of each column's elements equal to 1, as shown in 1.16.

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

Figure 1.14: A set of aligned sequences.
From [58], openly accessible for public use.

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}.$$

Figure 1.15: A Position Frequency Matrix corresponding to set of alignments above.
This Figure is from [58], adapted on October 2020, with permission of use for public.

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

Figure 1.16: The sum of all elements in a PPM is equal to 1.
The image is from [58], with open access for use on October 2020.

The table of probabilities or the Position Probability Matrix must be converted to a log form for even more efficiency. The final log scale matrix is referred to as PWM. To calculate the score of any potential binding site using PWM, each nucleotide's elements at columns of the matrix should be summed up[59]. Note that background frequencies can be calculated in a variant way. However, the most common way is to consider equal frequency for each nucleotide, resulting in ¼ = 0.25[57]. Then, each element of the PWM matrix is calculated using

$$PWM_{ij} = \log_2(\frac{PPM_{ij}}{Background_i})$$

as shown in Figure 1.17.

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.26 & 1.26 & -1.32 & -\infty & -\infty & 1.26 & 1.49 & -0.32 & -1.32 \\ -0.32 & -0.32 & -1.32 & -\infty & -\infty & -0.32 & -1.32 & -1.32 & -0.32 \\ -1.32 & -1.32 & 1.49 & 2.0 & -\infty & -1.32 & -1.32 & 1.0 & -1.32 \\ 0.68 & -1.32 & -1.32 & -\infty & 2.0 & -1.32 & -1.32 & -0.32 & 1.26 \end{bmatrix}.$$

Figure 1.17: The sum of all elements in a PPM is equal to 1.
The image is adopted from [58], on October 2020 with permission of access for public.

As it is observable in 1.17, applying logarithm may result in unacceptable values (infinities). To avoid such a scenario and eliminate null values, a sampling correction is added to each element of PFM before taking the log of each. This sampling correction is called "Pseudocount" and is calculated in variant ways regarding the tools and applications in use[59].

Sequence Logo

PWM is a vital and computationally efficient tool in bioinformatics. However, a more graphical version of it would give the researchers a better insight into the motif. "Sequence logo" is the graphical representation of PWM that can reflect a considerable amount of information to people in a visual approach. A Sequence Logo illustrates information content and relative frequency of a nucleotide at each consensus sequence position. This method would be easier to recognize the sequence motif and the most conserved base units at each position. There are computer tools available (refer to [60] and [61]) to construct a Sequence-Logo from a PWM or directly from aligned sequences[62].

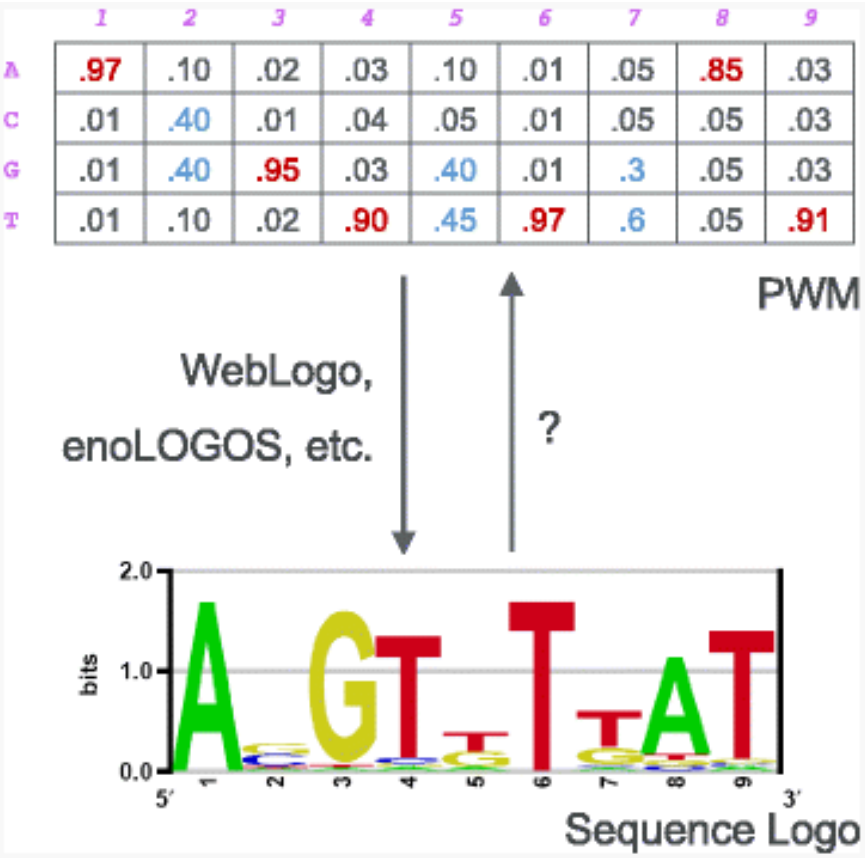


Figure 1.18: PWM can be directly employed to construct a Seq-Logo based on its data. However, recovering data for PWM from Sequence Logo is not straight forward.
Image is openly accessible on [62], and is adopted without any alterations on October 2020.

As described before, a PWM for DNA sequences has four rows, each for one of the nucleotides. The number of columns respects the length of aligned sequences, each associated with one of the positions at sequence. As shown in Figure 1.18, the sequence logo corresponding to the PWM on top has letters from different height and different colours. The colours are conventionally assigned for each nucleotide: green to A, blue to C, yellow to G and red to T. The height of each pile of letters at each position regards the Information Content of that column and is measured in "bits." Information Content is a measure to determine uncertainty. In DNA sequences, this

uncertainty can be phrased as how well-conserved a nucleotide is at a position[63]. Having P for PWM, Height for Information Content of each, i for each row, and j for each column (position), we can calculate the height of whole stack using the formula below :

$$Height_j = 2 + \sum_{1}^{4} PWM_{ij}. \log_2 PWM_{ij}$$

Regarding this, height of each letter is calculated using formula[62]:

$$Letter = PWM_{ij}.Height_j$$

Sequence Logos is an efficient way to transmit information about motifs. It contains various types of information that give the researcher a fundamental concept about the motif of study. First of all, letters (representing nucleotides) are arranged from the most frequent one to the least, from top to the bottom of the stack. So, the consensus can be formed by taking only the topmost letter of each position. In addition to this, the relative size of each letter at each position reflects how frequently that nucleotide has been observed at that position (The bigger the letter is, the higher is the frequency) and vice versa. Also, each pile's height at each of the positions corresponds to the amount of information in hand from that position so that the significant positions can be easily noticed. Besides, the sequence logo can be employed to demonstrate motifs within the aligned amino acid sequences. By all these being said, it should have been interpreted by now how important is the role of PWM and sequence logo in the field of bioinformatics and what a great value they carry in the case of computational analysis of biological data[64]. This makes these two modelling tools the core of many high-level bioinformatics studies to achieve more efficient and accurate results that help researchers gain a better insight into human beings' mysterious bio-mechanism.

Chapter 2

Related Works and In house Projects

To understand the research workflow and have a more profound insight into the bioinformatics behind my work, in this chapter, two projects proposed by our lab will go under consideration along with other related material. These projects are closely associated with the core subject and have been employed to develop and implement this thesis' main work.

2.1 TFBS Databases

2.1.1 JASPAR

JASPAR[65] is an openly accessible database of TF binding profiles (Accessible at `http://jaspar.genereg.net`), and one of the firsts of its kind. These profiles are stored as PFMs and TF Flexible Models (TFFMs) for TFs across variant species in six taxonomic groups. These predicted TFBSs could be accessed through the UCSC Genome Browser data hub[66] (access at `http://jaspar.genereg.net/genome-tracks/`)

that contains tracks for the human genome assemblies hg19 and hg38[65].

2.1.2 HOCOMOCO

Homo Sapiens COmprehensive MOdel COllection [67] is an openly accessible database of TFBSs that has been collected by integrating the data from both high and low-throughput techniques from available databases and reanalyzing them. This research's main goal was to develop a non-redundant database in which each TF contributes to the minimum possible number of models, while all models keep an acceptable TFBS recognition quality[67]. HOCOMOCO has been one of the data resources in this study and is accessible at `http://hocomoco11.autosome.ru/`.

2.2 TFBS Prediction Tools

2.2.1 TFBStools

One of the most coherent tools that have been developed for TFBSs analysis is called "TFBStools." TFBStools is an R package that provides the user with functions for matrix modification and DNA searching with those matrices' help. TFBStools facilitates data access and storage for the user with the help of well-designed S4 classes. Many valuable functions are developed to convert matrices to the desired form and compare a pair of them. It also makes it possible to search a DNA sequence for potential TFBS by scanning the sequence with Position Weight Matrices (see Figure 2.1). TFBStools also has a JASPAR database (an open-access database of TF binding profiles accessible at `http://jaspar.genereg.net/`) interface to present a better and more extant analysis to the user. It is also compatible with the two most common data formats in use for PWMs and generally provides an entire set of tools for TFBS analysis on a genome-wide scale[68].

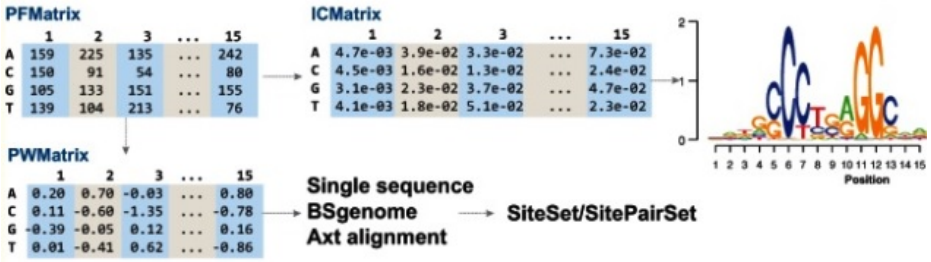


Figure 2.1: The general workflow of TFBStools(). Information Content and Position Weight Matrices can be generated from PFM. ICM can be used to make a sequence logo and PWM can scan single sequence alignments in order to exploit TFBSs which are stored in SiteSet object.
Image is taken from [68], an open access article, accessed on October 2020.

2.2.2 RSAT

Regulatory Sequence Analysis Tools or RSAT is a tool for analyzing cis-regulatory elements in the genome. This software suite provides the user with multiple functions for Motif Discovery in genome-wide scales, TFBS analysis (such as quality evaluation or comparison), comparative genomics, along with regulatory variations analysis. This software suite has been developed in a modular manner and can be employed individually and as part of a pipeline to perform high-level tasks. This tool has a web interface that is accessible for users at `http://rsat.sb-roscoff.fr`[69].

2.3 In House Projects

2.3.1 MethMotif

If we respect the classic definition of genetics, it is impossible to justify a number of events such as phenotypic diversity in a population, the different phenotypes observed between monozygotic twins, or their different susceptibility to variant disease notwithstanding that they share the same DNA. So, to be able to explain these phe-

nomena, scientists look at them from a different aspect named "Epigenetics"[70]. This field was first proposed in 1939 under definition: "the causal interactions between genes and their products, which bring the phenotype into being[71]." This definition then was taken over with: a set of heritable variations in gene expression that are not caused by modification of DNA sequence. Among all epigenetic markers, DNA methylation is the most significant one. Many remarkable discoveries were done focusing on DNA methylation and representing how it alters gene regulation, which resulted in the development of human epigenome projects and epigenetic therapies. One of these projects is called "MethMotif".

DNA methylation has been reported as an influential factor in attracting TFs and stressing the necessity to connect TFBSs with their associated DNA methylation profile. MethMotif is a two-dimensional TFBS database that studies the Position Weight Matrix of a TFBS accompanied by cell type-specific CpG methylation information. "The CpG sites are regions of DNA where a guanine nucleotide follows a cytosine nucleotide in the linear sequence of bases along its 5' → 3' direction. CpG sites occur with high frequency in genomic regions called CpG islands (or CG islands)[72]." (See Figure 2.2.)

Figure 2.2: A C-Phosphate-G (CpG) site from 5' to 3' is indicated as yellow.
The image is from [72] with permission of use for public, on October 2020.

MethMotif combines information driven from ChIP-seq and Whole Genome Bisulfite Sequencing (WGBS) for better characterization the location on DNA that was

targeted by TF. The ChIP-seq and WGBS data were obtained from ENCODE consortium and The gene expression Omnibus[73] (complemented by in house generated data) databases[27]. After integrating ChIP-seq and WGBS data-sets, MethMotif classifies the DNA-interacting proteins regarding methylation profiles of the DNA. This integration gives the power to MethMotif to profile DNA methylation landscapes surrounding binding locations of DNA-bound proteins at a genomic scale. In addition to this classification, MethMotif presents a new two-dimensional representation by combining the Position Weight Matrix of a TFBS and DNA methylation (refer to Figure 2.3). This representation is very insightful because it is a more expressive way to reflect DNA methylation's impact on recruiting the TF of interest. To explore MethMotif's data-sets and to easy access to its database, a web-based interface has been developed which is reachable at `https://bioinfo-csi.nus.edu.sg/`. The access to MethMotif is available through the following three modes[27]:

1. 'Motif database direct query': Using the proper ID defined by MethMotif or official gene name, the user can look for all DNA binding proteins stored in the database.

2. 'Explore': Users can intuitively investigate DNA binding proteins. Embedded dynamic heat-maps represent each protein categorized based on its CpG methylation pattern, with a range of 200 base pairs surrounding ChIP-seq's peak summit.

3. 'Batch query': MethMotif provides the user with tools to facilitate the binding sites' characterization targeted by a protein of interest and other co-factors. This observation is obtained by analyzing the existence of TFBSs and their Methylation condition across a list of genome loci.

At the time of writing this thesis, MethMotif database documents over 655 TFBSs computed from over 2473 ChIP-seq data-sets in 16 different cell types[74].

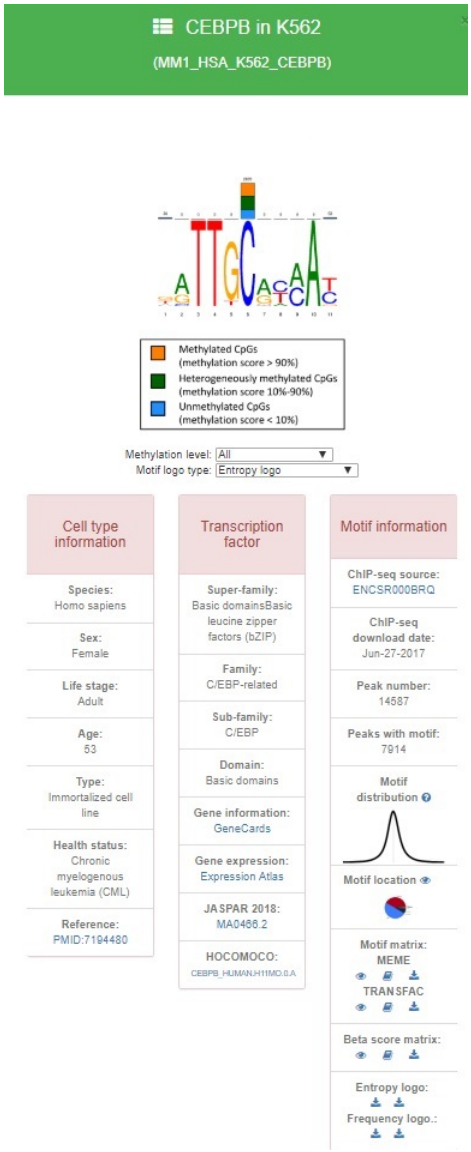


Figure 2.3: Available data in MethMotif's website, for TF CEBPB in context of bone marrow cell line (K562). MethMotif covers two most well-known set of data formats. It provides information about TF, cell type and motif.
This image is adopted with alteration from [74], with the permission of the author, accessed on October 2019.

2.3.2 TFregulomeR

TFs target the DNA sequence through a highly dynamic procedure that includes a vast set of DNA segments. To increase the study's resolution in this matter, it would be quite advantageous if the broad range of the study narrows down by studying TFs in a particular condition. Analyzing TFs in cell-specific context is the approach taken in one of the hosting lab's projects named "TFregulomeR," which is an R-package containing variant functions to help people manipulate and analyze TFBS and methylome meta-data. TFregulomeR derives data from MethMotif and Gene Transcription Regulation Database (GTRD)[75]. This library is especially useful to characterize TFs that work as partners to bind to the DNA sequence and analyze them in different partnering cases along with DNA methylation level data.

TFregulomeR's extensive TFBS database can be explored by a simple function (TFBSBrowser) with the option to choose specific cell/tissue type, sample type, organ, or species by providing the arguments to narrow down the search space.

As it was described in previous sections, the main goal of a ChIP-seq experiment is to locate protein binding regions or "peaks." The raw output of the ChIP-seq experiment is a large set of short DNA fragments referred to as "'reads." Large piles of these short fragments clustered at a region after aligning the whole set of reads form peaks, which can correspond to a binding site. By daily decrease in sequencing cost and the development of new technologies, it has become possible to compare data derived from different ChIP-seq experiments, for instance, to examine the binding of a protein of interest in a specific condition such as the presence of another factor. One of these comparative methods, which is commonly used, is called "Overlapping Analysis." In this method, peaks called by different methods are compared and categorized into "Common peaks" or "Exclusive Peaks" in such manners that those regions that are called as peaks in both experiments are called "Common Peaks." Note that the pro-

portion of common peaks from each set corresponds to the qualitative amount of that very set compared to the other one, also to a predefined threshold which is set by experiment[76]. TFregulomeR() package provides the user with a valuable function to study the co-binding landscape of two sets of TFs. intersectPeakMatrix() has been designed to study co-factors of TF of interest in cell-specific context by taking two sets of peaks either from the user or TFregulomeR's database. This package is accessible for everyone at `https://github.com/benoukraflab/TFregulomeR`.

Chapter 3

Research Question and Results

TF's remarkably dynamic binding performance is an influential reason to study the TFBSs in a cell-specific context. One problem with current methods is that contemporary databases represent TFBSs studied in several cell lines combined (and not in a cell-specific way). Secondly, they rely on ChIP-Seq assay's outcomes, even in the case of dimer complexes. This is done based on the assumption that TFs bind to the DNA sequence independently, resulting in an inadequate depiction of dimers. Thirdly, these databases have not taken the impact of the methylation profile of a TFBS into account, which is reported to have an undeniable effect on the binding preference of TFs. Though there are differing DNA targeting approaches for TF families, the modelling methods for predicting the target sequence are the same for all. In addition to that, the current PWM based pattern matching models that are employed for motif prediction suffer from some drawbacks, although they are highly used for their simplicity. Besides, it is reported that around 40% of almost 1,400 sequence-specific TFs encoded in the human genome are not characterized yet[77]. The abovementioned are the problems we are tackling in here, and for the explained reasons there is a high demand for research and development in this particular field. In this chap-

ter, the study's question will be explained further, and it will be described how we approached these issues.

3.1 Question of Study and Current Limitation

Several factors should be taken into account while studying TFs' behaviour in targeting a specific DNA sequence. That is important because there are features other than TF-DNA sequence interactions that can alter the DNA sequence being targeted by the TF (refer to Figure 3.1). Some of these features are the impact of co-factors that bind to TF of interest, DNA modifications to interplays between a TF of interest with another TF[77].

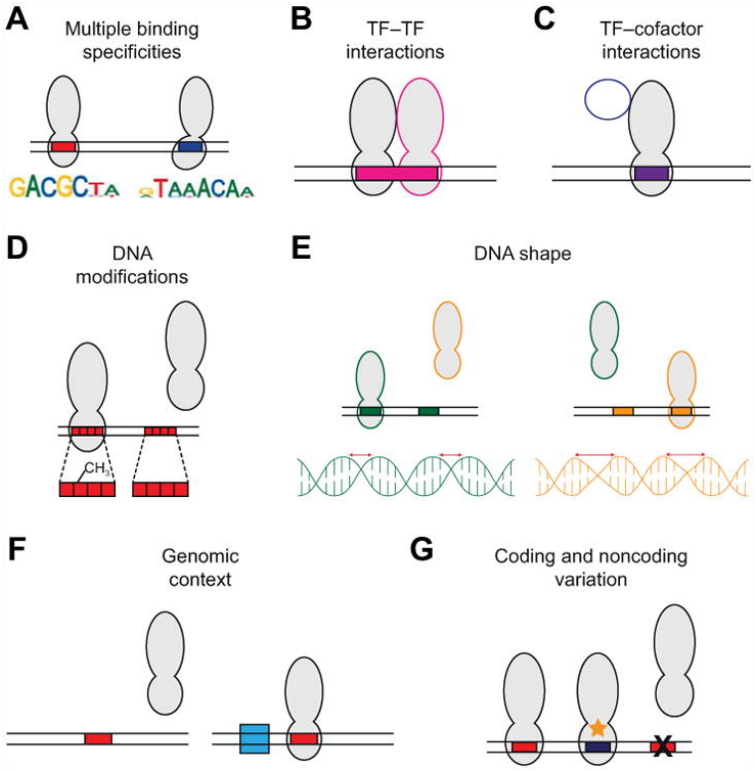


Figure 3.1: Other than binding specifications that regulate gene expression by attracting TFs, other features affect this phenomenon. TFs can interact with each other (B) or other structures present in the environment to alter their behaviour (C). DNA modification (such as methylation) is an influential event (D) too. In addition to these, DNA's special structure can make TFBSs accessible for TFs or vice versa. Also, some modules distant from the binding site can affect the area owning to DNA shape(F). Also, the variation in the gene itself is one of the possible scenarios (G). *This Figure is adapted from [77], with the permission granted from the Elsevier publication, on October 2020.*

The whole set of TFs is categorized into over 60 families regarding the structural similarities of them. One of the largest families of TFs is called "Basic Leucine Zipper" or bZip[78].

We are focusing on this family for development of our project. This family of TFs regulates the expression of many genes, and they can bind to over one hundred variant DNA that corresponds to specific disorders, which makes them a hot topic in studying TFs. Still, the question is why this family has such a broad sequence preference?

The answer lies behind the structure of this set of TFs and how they work to target DNA sequence. bZip TFs work in groups of two to target DNA sequence (see Figure 3.2). Basically, they first pair up with a molecule similar to themselves (forming a homo-dimer) or with a distinct TF (constructing a heterodimer), then they collaboratively target a sequence within the DNA and bind to it. The change in the partner of each TF alters the sequence preference. Although there are several cases that the mechanism of this group work has been studied, it is not yet discovered how the bZip dimer targets a fragment of DNA. In many cases, the target sequence of bZip dimer is related to the individual target of one of the partners. However, numerous cases show that the DNA preference of bZip dimers cannot be predicted based on each of the TFs' independent preferences. This change in the behaviour of bZip TF in the presence of a different partner expands the number of their target DNA sequences and makes them a compelling case of study for scientists being a highly dynamic set of TFs[79].

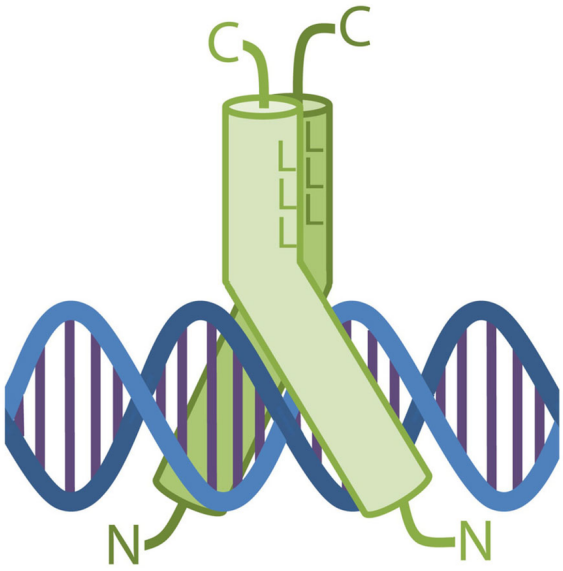


Figure 3.2: bZIP dimer binding DNA. The L is for the leucine forming the interface in the bZIP dimer.
This image is adopted from [80], with open access for public on October 2020.

The popular PWM based pattern matching models that are employed for motif prediction suffer from several drawbacks, although they are highly used for their integrity. First of all, they are susceptive to the quality and quantity of the DNA fragments on which the matrix is built. Also, a high rate of false-positive results has been reported in studies that use PWM. In addition to these, the current modellings never depict the relation between independent positions of TFBSs. Besides, to overcome the individual model's limitations, variant models are constructed for a single TFBS to reflect sequences' variation. Finally, PWM models do not reflect any information about TFs' alternate structure, neither the methylation profile.[54].

As depicted in Figure 3.3, the conventional model representing a TF of interest for all TFs is of the form of the top captioned logo as the "global motif." This logo represents the set of sequences that have been targeted when one member of the bZip family, named CEBPB, was studied in different cell contexts. Many of sixteen ChIP-seq experiments led to forming sequence logos similar to "Global Motif" constructed from

two parts: one half represented a CAAT (or its complementary sequence reversed: ATTG), and the other half a mixture of ATTG/CAAT and TCA/TGA. The reflection of ATTG/CAAT, a famous binding site for CEBPB in one half of the logo and the ATTG/CAAT and TCA/TGA mixture on the other half, was a piece of evidence on the fact that CEBPB is working with identical or dissimilar TFs.

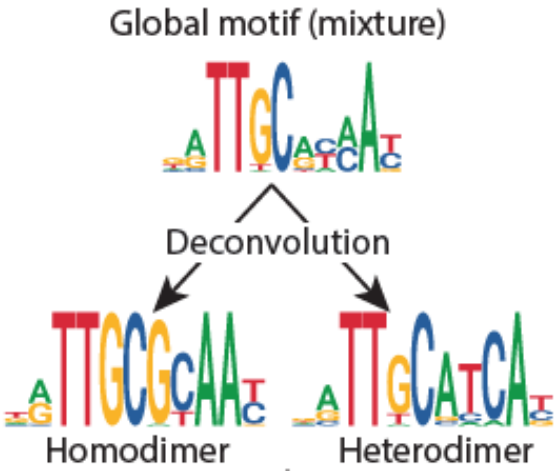


Figure 3.3: studying bZIP dimers for motif discovery, results in a matrix which is the mixture of data corresponding to homodimer and heterodimers. The global motif belongs to the CEBPB. The exclusive impact of TF of study (CEBPB in here) is obvious in left half of each logo.
The Figure is adopted from [81] with modifications applied with the permission of the Oxford University Press publications, accessed on October 2020.

Based on the abovementioned, the problems can be narrowed down into the following: First, current high cited databases, such as JASPAR or MethMotif, are still using the motif of CEBPB and a mixture of co-factors as CEBPB's binding site motif, as experiments like ChIP-Seq captures the entire data of both homodimers and heterodimers. In addition to that, current modelling methods are not expressive enough, because not only merging all the collected data in order to form a single sequence-logo makes the final Figure a roughly precise model, and the matrix of TF a systematically noisy one, but also a single sequence logo does not reflect any information about TF's

structure or contribution of other co-factors. We have proposed a novel model to tackle this issue, which will be described in the following section.

3.2 Proposed Model, Methodology and Results

As was discussed in the last chapter, TFs of bZIP family work in pairs to target DNA sequences, and variation in the partner can alter the sequence preference. However, the homodimers and heterodimers are captured together by assays like ChIP-seq, resulting in a noisy PWM and subsequently a Sequence-logo of two halves: a conserved and a degenerated half. For example, take the sequence ATTGCGCAAT captured from a homodimer complex, and ATTGCATCA is the heterodimer's binding site. The resulting motif for a big set of these sequences may be ATTGCACCAT, which is not an existing sequence on DNA (see 3.3 for better insight). To deal with such false positives, we proposed a model named "Forked Position Weight Matrix" or FPWM, an R library for providing the user with a better description and representation of TFs precise characterization of TF dimers. FPWM is a PWM that systematically keeps the conserved half of the traditional PWM as data of TF of interest and specifies the region effected by co-factors by forking the matrix multiple submatrices. The visual representation of FPWM would be a graph of sequence logos, with one parent node representing the binding site of TF of interest, and two or more leafs representing each of the co-factors under study. This form of visual representation for a TF and its co-factors is more expressive, but the novel data format employed for this purpose is compatible with all the well-known tools. In the following sections, the method and material will be described.

3.2.1 Data flow

The Forked Position weight matrices generated by the FPWM package are obtained from the MethMotif, which records data compiled from MethMotif's own data sets and GTRD. As discussed, MethMotif integrates TF motif data with the methylation information, representing a two-dimensional sequence logo of a motif accompanied by associated methylation information. On the other hand, GTRD has been an external resource to access ChIP-seq peaks called MACS peak caller. Each PWM has its unique ID representing its resource (MethMotif or GTRD), species, cell type, and TF's name. The current version of FPWM works with the collection of peak sets available through TFregulomeR. FPWM exports an intersection matrix with an indication of the intersected peak percentage for each pair of peak lists selected by the user. This peak intersection matrix, complemented with DNA methylation status, is a valuable tool for co-factor and TF interaction analysis. The FPWM complies multiple intersection peak matrices by receiving the ID of TF of interest and its co-factors that the user is willing to study. The peak list of TF of interest will be the first peak list in all intersection matrices. The second peak list for each intersection matrix will be for one of the indicated co-factors. FPWM provides the users with functions to help them choose the most significant co-factors, plot the graph, or generate the Forked Position Weight Matrix. More details can be observed in Figure 3.4.

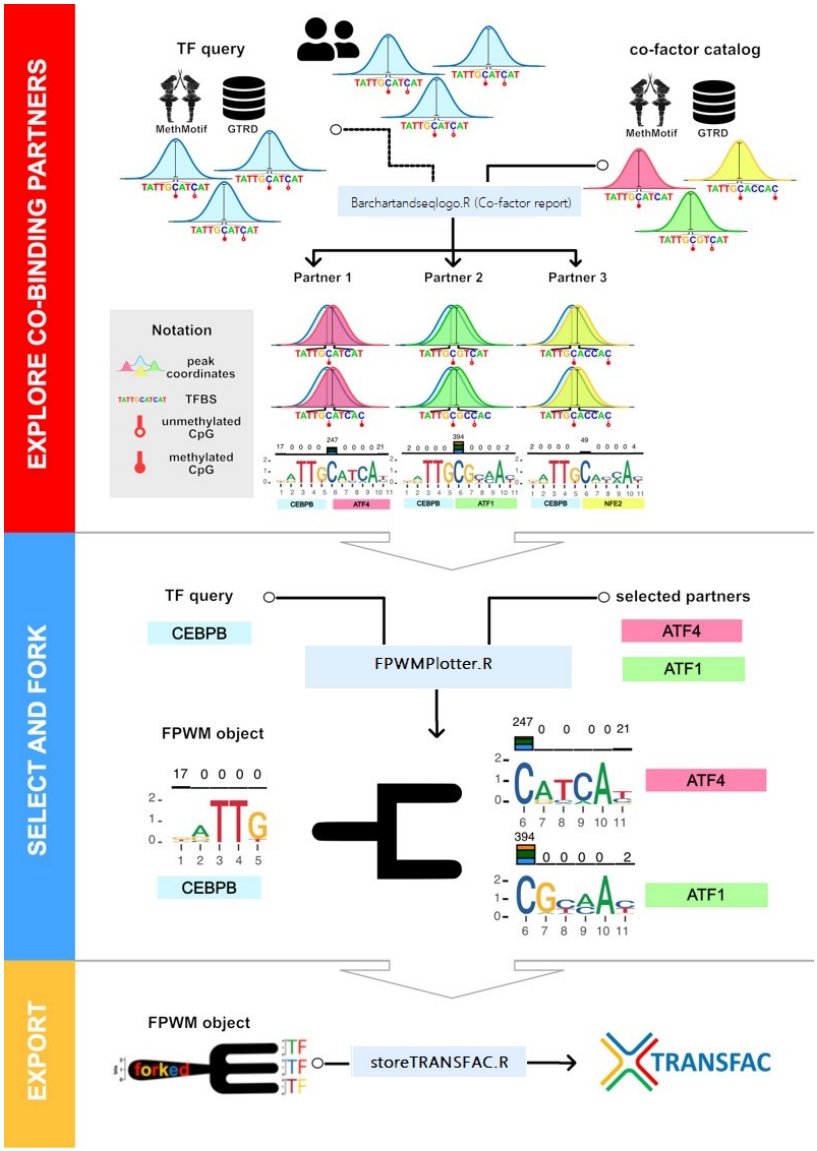


Figure 3.4: This diagram depicts the flow of data and how FPWM accesses external sources and tools. In the upper level, exploring the co-binding partners with embedded functions is explored to describe top TF binding partners by overlapping the query coordinates, based on co-binding percentage or enrichment score. This step is followed by selecting a forking position and creating an FPWM based on the number of top co-binding partners. In this level, individual matrices are generated regarding the intersection of binding partners. In more detail, the main TF peaks are segregated into individual PWMs, which are created only from intersecting peaks with top co-binding partners. Eventually (but not limited to), the resulting deconvoluted matrices can be exported as a TRANSFAC file employed by other programs such as RSAT etc.

The figures is generated collaboratively by members of lab and with the contribution of Walters Santas.

3.2.2 Functionalities

FPWM provides the user with multiple functions in order to ease data visualization and analysis. These functions can obtain and store intersected matrix of several peak lists of cofactors from TFregulomeR(), modify them to generate FPWM, and store them in an organized manner as an S4 class object. Some internal functions are embedded to merge the first half of multiple PWMs and generate a proper parent matrix of TF of interest and methylation information. FPWM library can store the local file of a novel data format for multiple proposes. It also provides the user with the graphical list of cofactors of TF of interest in the order of their significance. Eventually, the FPWM library can read and analyze locally provided files and generate a graph of the sequence logos for a TF and its cofactors with an optional methylation level plot on top of them. All these functionalities are implemented in an R-library package.

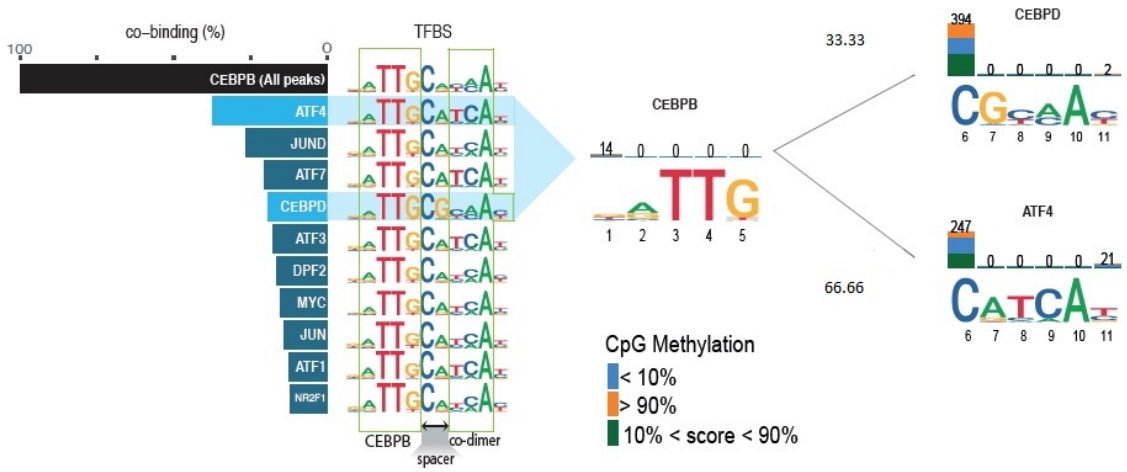


Figure 3.5: The bar chart on the left is the outcome of one of FPWM's functions. Barandseqlogo() function receives the name of TF of interest (CEBPB in this case) and plots a chart containing all the TF's cofactors that are accessible via TFregulomeR(). Each bar represents the co-binding percentage of TF of interest with that cofactor, based on each pair's intersected matrix retrieved from TFregulomeR(). Each Sequence logo next to each bar represents the PWM of it. By this vision, two cofactors have been selected for the FPWM plot. Note the difference between the spacer region of homodimer (CEBPD) and heterodimer (ATF4). On the right side of the figure, a Forked-PWM is depicted for two cofactors of CEBPB, using the FPWMplotter() function. The weights on edges are relative co-binding percentages of each cofactor. In this case, methylation levels have been plotted too.

3.2.3 R Object for FPWM

The Forked Position Weight Matrix works with an object in order to efficiently retrieve and store data. An embedded function generates an s4 class object with multiple slots for intersected storing PWMs derived from TFregulomeR. In addition to that, IDs, forking position, overlapping percentages, and the novel Forked Position Weight Matrix, along with methylation information for each profile, are stored in one class. This approach helps users to access data more straightforwardly. A simplified scheme of the FPWM library is provided in the Figure 3.6.

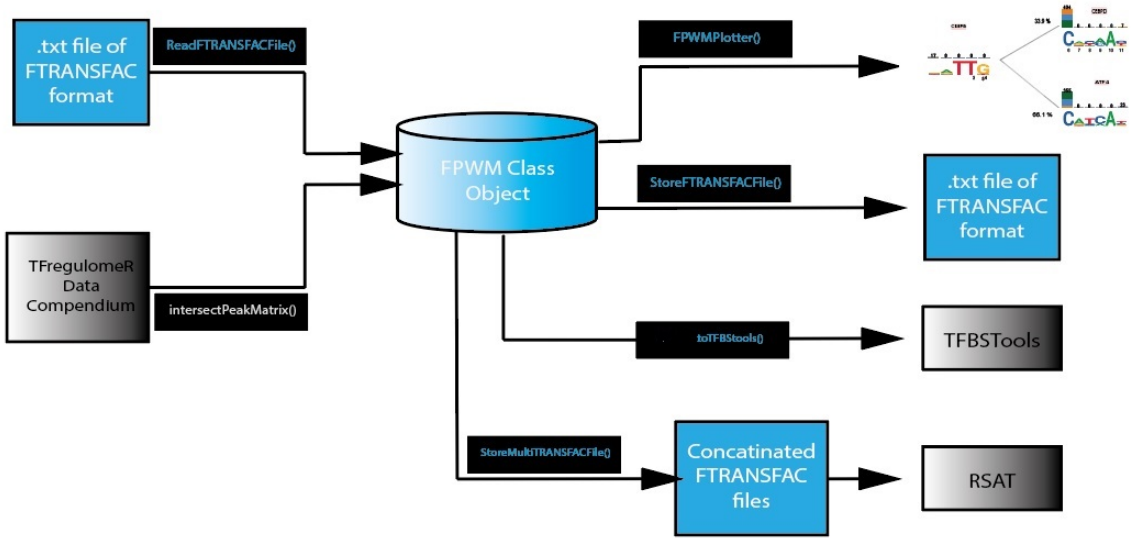


Figure 3.6: A S4 class is created once the package is installed to store user-provided data along with the content derived from TFregulomeR(). This class-oriented approach makes it easy to modify and access data for users and developers. As it is depicted, multiple functions have been embedded in the plot or generate files. FPWM's outcome can be received using RSAT or TFBStools() for genome-wide analysis.

3.2.4 Novel Data Format

A typical TRANSFAC format of data is depicted in Figure 3.7.

```
AC MM1_HSA_K562_CEBPB
XX
ID CEBPB
XX
DE MM1_HSA_K562_CEBPB CEBPB ; from MethMotif
PO  A   C   G   T
1   93  61  192 187
2   305 74  141 13
3   0   0   0   533
4   0   0   0   533
5   49  0   390 94
6   0   533 0   0
7   280 59  139 55
8   91  205 3   234
9   290 242 0   1
10  533 0   0   0
11  15  215 73  230
XX
CC program: MethMotif
XX
//
```

Figure 3.7: A PFM of CEBPB's motif, organized in TRANSFAC format, derived from MethMotif[27]. First half shows the ID and some additional information, the second half, holds the PWM matrix with an extra "Positions" column named PO representing each position in a sequence.The content terminates with a XX followed by //.

FPWM derives data on intersected matrices from TFregulomeR() (see left side of Figure 3.8), then regarding a user-defined forking point, merges the conserved region of global motif into one single matrix (parent), and concatenates the degenerated part of each intersection matrix to the parent, forming a single profile matrix for them all which is called Forked Position Weight Matrix (refer to right side of the Figure 3.8).

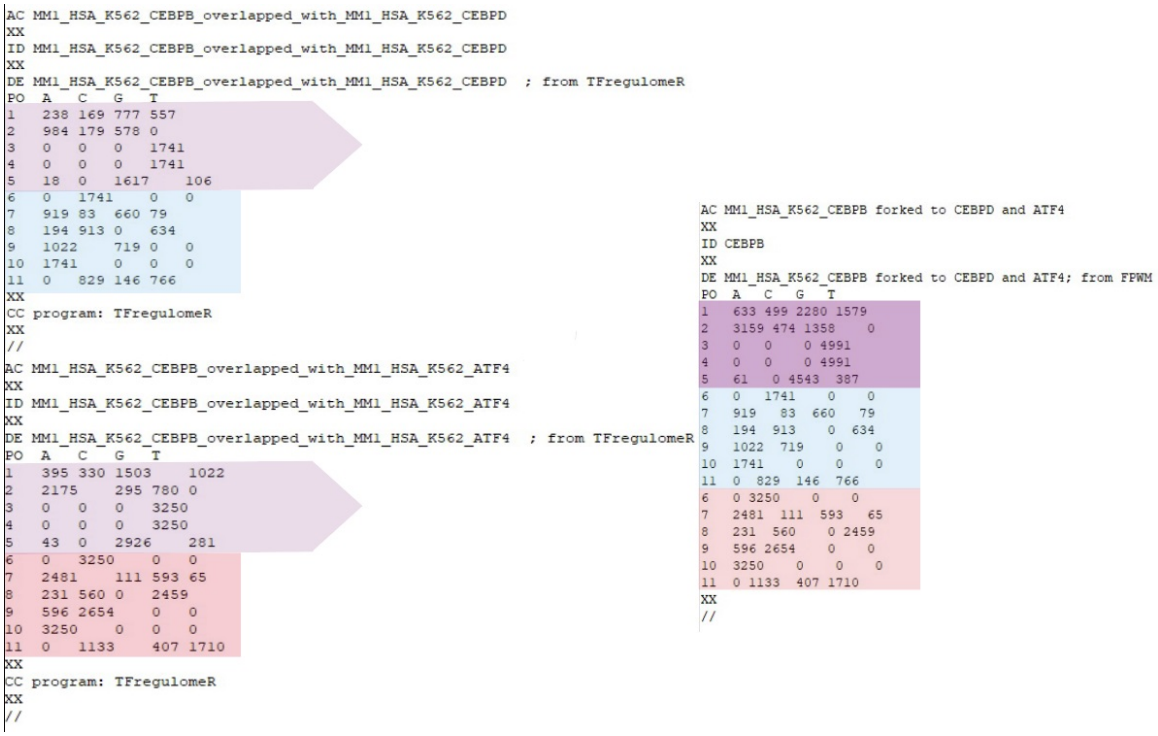


Figure 3.8: A simplified scheme of how FPWM is constructed from only two matrices. Based on a forking points (5 in this scheme), two matrices are splitted and merged up to the Forking point. The second half of matrices follow the merged matrix immediately. Notice the repetition in positions, than indicate forking position.

After construction of the FPWM with this approach, FPWM storing functions can degenerate the data files of TRANSFAC format as shown in Figure 3.9.

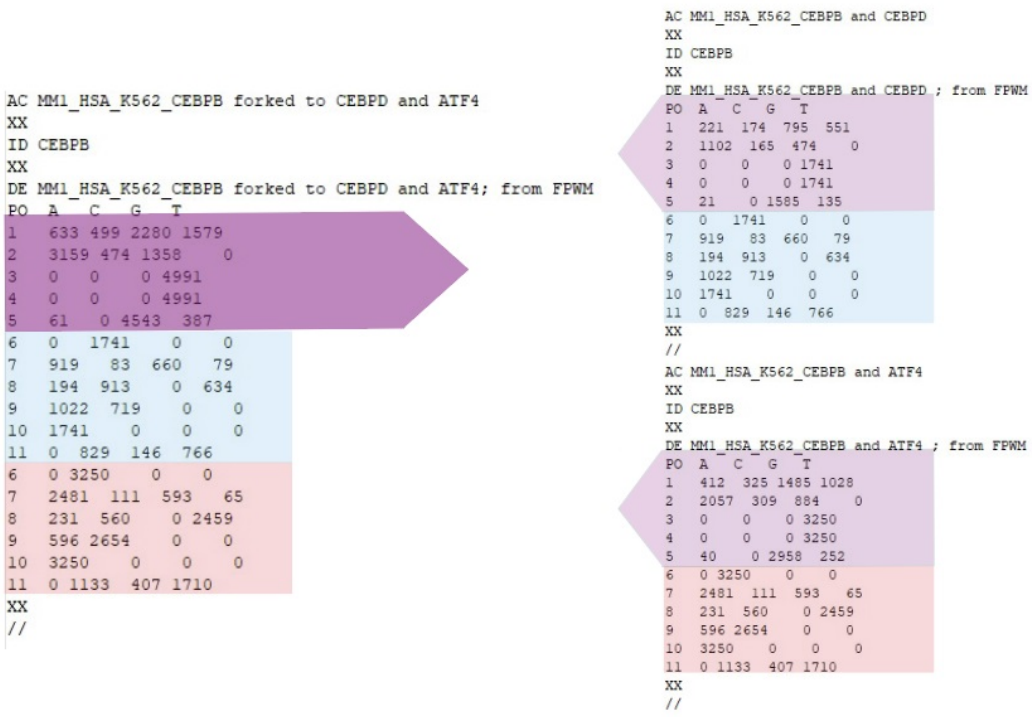


Figure 3.9: A file of Forked-TRANSFAC format, degenerated into two regular TRANSFAC formats, for CEBPB overlapped with CEBPD and ATF4. The merged section of each FPWM, holds the total number of CEBPB peaks overlapped with ATF4 and CEBPD (4991 in here). In order to form new matrices, the elements in merged part, are divided by number of peaks at each of the subsequent co-factors. (1741 for CEBPD and CEBPD and 3250 for CEBPB and ATF4)

With the help of this new format, it is possible to depict all possible co-factors of a TF of interest in a single file, along with some other implicit information such as the total number of TF of interest's binding sites with its co-factors, and the number of each binding sites for the TF of interest and each specific co-factor. To generate the FPWM, the first half of all matrices are added up (resulting in the total number of overlapped peaks of TF of interest with its co-factors). Then, for forking them into each of matrices, the merged matrix is divided by the co-factor matrix's value. This value is calculated by dividing the total number of overlapped peaks between TF of interest and all its co-factors by the number of overlapped peaks between TF of interest and the target co-factor. For example, regarding the Venn diagram in 3.10,

the value for the CEBPB-CEBPD matrix's element will be 4991/1741. Note that in all steps, numbers are rounded to integers to respect the TRANSFAC format.

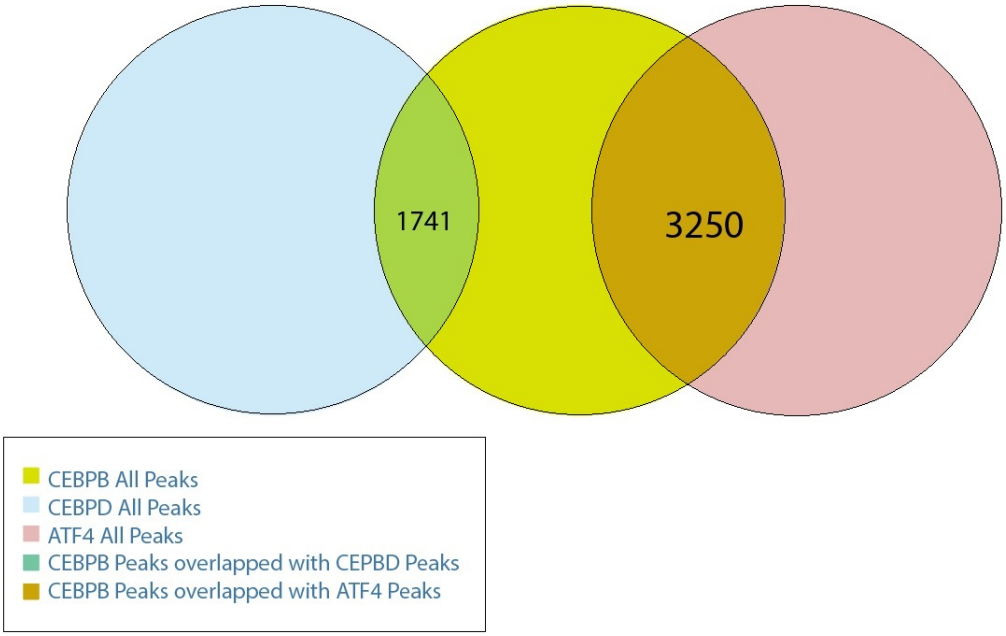


Figure 3.10: The number of overlapped peaks TF of interest (CEBPB) and two of its co-factors (CEBPD and ATF4). We assume the amber circle in the center represents all the set of peaks determined for CEBPB, and similar to that, the pink circle representing the peaks of ATF4 and the blue one for CEBPD. As can be seen, these sets share several peaks, which would be how overlapping peaks are determined. Note that this Figure is simplified for a better understanding. There can be scenarios in which all the circles have shared regions.

This approach to store data, inspired by TRANSFAC data format, allows us to organize and keep data in an efficient way and is a more expressing approach to depict the presence of co-factors and the level of their involvement in our case of study. Using this method, the generated matrices are tested and evaluated in the next section to discuss how these forked matrices perform better than the resource database's: MethMotif.

3.2.5 Evaluation and Results

The main goal of developing the FPWM package is to highlight the neglected issue that dimers have more dynamic sequence preferences. To evaluate our proposed model, we need to investigate its power in predicting the dimer complexes' binding site. We aim to scan the binding site of these dimer complexes with FPWM and then compare it to the global matrix of the target TF. Our evaluation shows that FPWM shows a higher match score than the current motif profiles, thus improving the TFBS prediction power.

The FPWM evaluation has two distinguished phases. The first phase is to see whether FPWM for a target TF and one of its co-factors perform better in the sense of predicting the dimer's binding site, compared to the global matrix of target TF. The second phase would be evaluating FPWM in the context of different cell lines. In the later one, we aim to show that cell-specific FPWM performs better compared to a general matrix profile for targeted TF. Thus, the evaluation steps can be seen as below:

1. Targeting one TF and a set of its co-factors in one cell-line.

2. Targeting one cell-line and a set of other cell-lines for one TF that exists on all of them.

At each of these steps, FPWM will be scanning the set of sequences that are common between target TF (cell-line) and one of its co-factors (cell-lines). The Venn diagram below, is for the clarification of this fact.

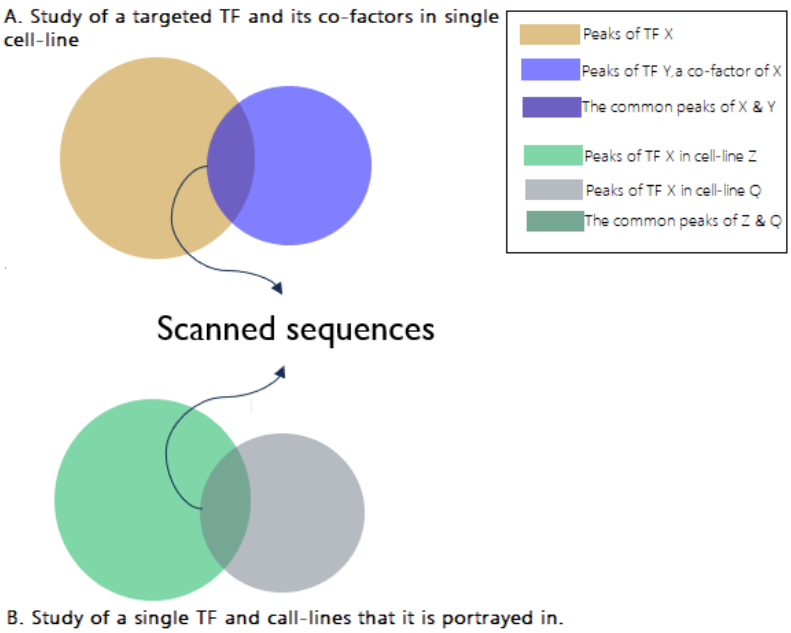


Figure 3.11: Scanned sequences: there can be two possible scenarios for the set of peaks that are scanned using the related matrices. In scenario A, we are studying a given TF and its co-factor. As explained, these two TFs share a subset of their peaks representing those regions in which they form a dimer. These regions should be closely studied to show how FPWM, designed for dimer complexes, outperforms the general matrices. In scenario B, we are trying to evaluate the cell-line specific behaviour of FPWMs. Here, we target one TF then trying to study FPWM's performance on a combination of several cell-lines for the same TF. For this purpose, the set of sequences that are a TFBS for the given TF shared among multiple cell-lines are scanned. We aim to see a more powerful TFBS prediction by excluding noises coming from the integration of extra cell-lines.

The set of peaks are exported as a .BED file. "BED (Browser Extensible Data) format provides a flexible way to define the data lines that are displayed in an annotation track[82]." This file holds information about genomic intervals or in this context chromosomal location ranges. By convention, the locations are written in the notation of Start $<$ End for both strands of DNA; thus there is just one coordinate system for this matter.

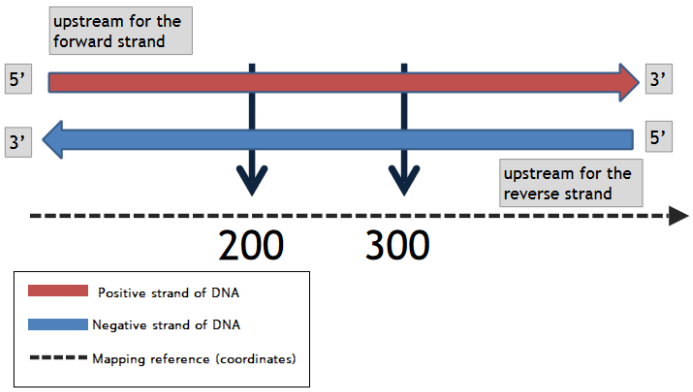


Figure 3.12: An example of a genomic interval and how the coordinate system works, adapted from [83] with the permission of author, with minor modifications. As mentioned, we are dealing with two strands: Positive and negative in locating a specific region on the DNA strand. Although these are opposite strands, one coordinate system is used to navigate the genome. In the Sequence Alignment Map, a binary flag indicates the strand being negative or positive.

As can be seen in Figure 3.12, each strand has two directions. The upstream region is considered the region before 5' and relative to the direction of the transcription. With all these in mind, a BED file is a file including 3 requires columns: Chromosome name, Start and ending coordinates. Note that this file may contain 9 and more optional columns, respecting the information that is needed for the study. An example of a simple .BED file can be seen in Figure 3.13.

```
chr1 213941196 213942363
chr1 213942363 213943530
chr1 213943530 213944697
chr2 158364697 158365864
chr2 158365864 158367031
chr3 127477031 127478198
chr3 127478198 127479365
chr3 127479365 127480532
chr3 127480532 127481699
```

Figure 3.13: example of a simple BED file with only required fields

With all the above mentioned in mind, at each stage, we will study the performance of the FPWM compared to the global matrix. The evaluation methods are shown in the number of cases in the following subsections.

FPWM for different co-factors of a targeted TF

In this subsection we take a look at different co-factors that one TF interacts with. Two cases are presented in here. In first one the target TF is CEBPB and the second one is MAFF.

Case 1 : CEBPB and its co-factors in K562

The process of constructing FPWM for a given TF's co-factor, starts with taking a look at the quick report of main co-factors of CEBPB, ranked based on the size of common peaks.

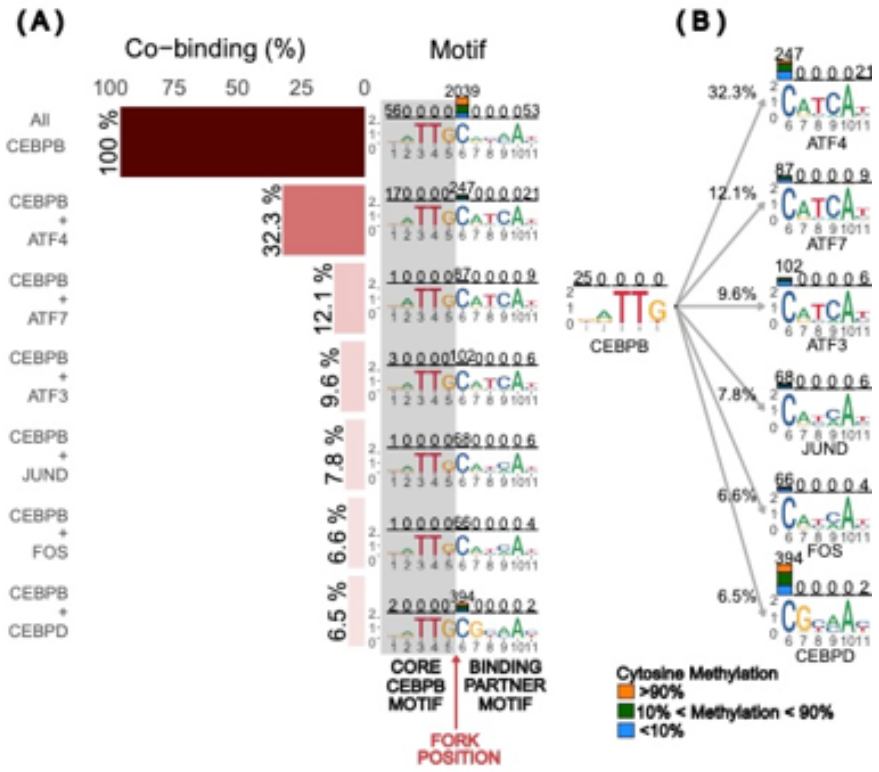


Figure 3.14: CEBPB co-factor report: Using the functionalities embedded in the FPWM package, the user can start by taking a look at the most significant co-factors of a given TF. As shown in the figure, the report is composed of two main sections: the bars on the left side, are representing the amount of overlapping that occurs between peaks of two TFs (names are shown on the left side of the bar.) On the right side, the motif profile for the set of overlapping peaks is represented, coupled with the methylation profile of the same TFBS. This helps the user detect the conserved region (highlighted in gray), which corresponds to the given TF (CEBPB in here) and subsequently setting a proper forking point.

By a quick look at the report, it can be easily noticed that from position 5 on, the matrices are starting to differ. Note that each of these matrices is constructed for the intersected peaks between TF and its co-factor. Thus, the conserved region (in the left) and different sequence preference of each CEBPB's partners is easily noticeable in the right half of the sequence logo. Note that in this report, not only the DNA sequence profile matters but also does the sequence methylation profile.

With the interpretation coming from the co-factor report, we can move on and con-

struct an FPWM for the co-factors seen in the report.

As a result, an FPWM for CEBPB, along with all the specified co-factors, will be constructed. This matrix can be segregated into exclusive matrices for each of the dimers then used for TFBS prediction. In the following, we used the CEBPB-ATF4 matrix to scan the common peaks coming from the ChIP-Seq assay of each of those TFs. As seen in Figure 3.15, it outperforms the global matrix for CEBPB.

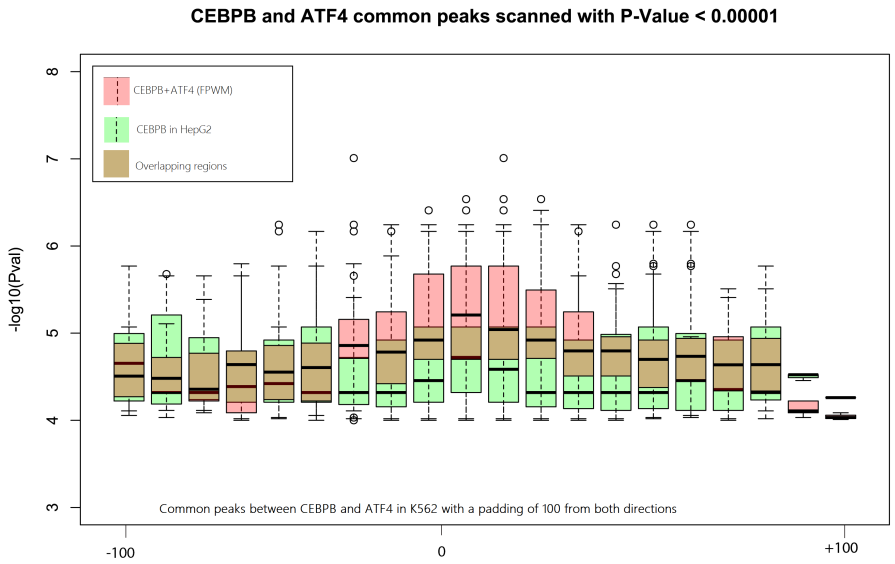


Figure 3.15: CEBPB-ATF4 FPWM performance analysis on the common peaks of the CEBPB and ATF4 in K562: In this bin-map plot, we compare the performance of two different matrices. An FPWM of CEBPB+ATF4 (shown in pink) and the global matrix of CEBPB (shown in green) are used to scan the set of peaks that CEBPB and ATF4 are sharing (which corresponds to the CEBPB-ATF4 heterodimer). The summit of peaks is in the center of the graph, and they have a padding of 100 bases from each direction. As it is depicted, the FPWM shows a better match (higher P-value) as the scanner approaches towards the TFBS peak summit. This concludes better prediction power of FPWM compared to a general matrix. Note that both matrices are in cell0line K562.

Case 2 : CEBPB and its co-factors in HepG2

In order to showcase this applies for all the cases, we did the same procedure for a different cell line. We construct FPWM this time with passing the argument cell=HepG2. As can be seen in 3.16, our FPWM for CEBPB-ATF4 dimer, in HepG2 cell line, shows better match thus higher prediction power compared to global matrix for CEBPB.

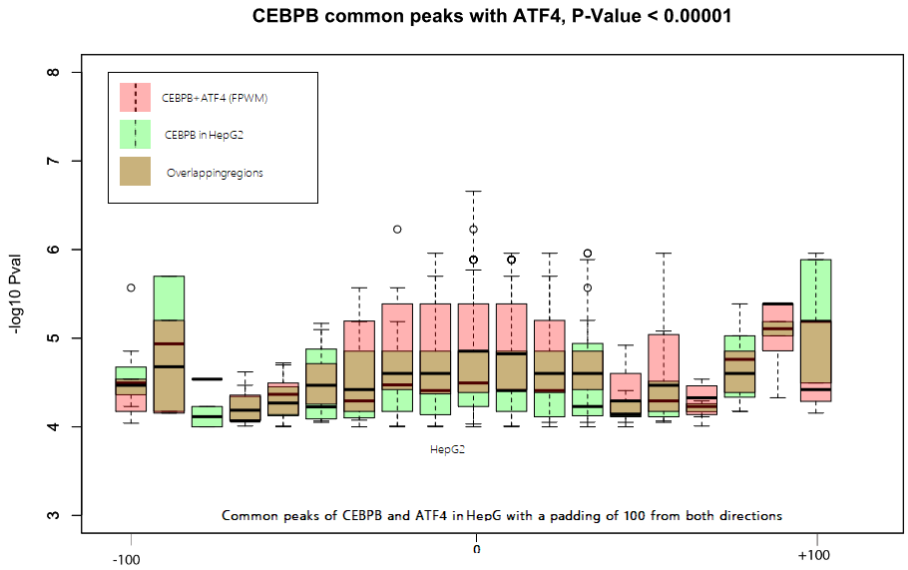


Figure 3.16: CEBPB-ATF4 FPWM performance analysis on the common peaks of the CEBPB and ATF4 in HepG2: Similar to the previous case, in here we are comparing the performance of FPWM of CEBPB+ATF4 (shown in pink) and the global matrix of CEBPB (shown in green), which are used to scan the set of peaks that CEBPB and ATF4 are sharing only this time in the other cell-line (HepG2). The summit of peaks is in the center of the graph, and they have a padding of 100 bases from each direction. As it is depicted, the FPWM shows a better match (higher P-value) as the scanner approaches towards the center of TFBS. This shows the better prediction power of FPWM compared to a general matrix and proves how our model works properly in all selected cell-lines.

These examples are two of many only to show that performance evaluation is not restricted to one cell line and our tool works well for all the scenarios. To extend this claim, we are going to study another dimer in the following subsections.

Case 3 : MAFF and its co-factors in K562

In order to observe other set of dimers in the cell line K562, we targeted MAFF as out main TF. After going through the analogous procedure to construct a FPWM, we use the matrix for MAFF-MAFG dimer to scan their intersected peaks. As it is evident in 3.17, the FPWM for MAFF-MAFG dimer, works better compared to global matrix for MAFF that exists out there.

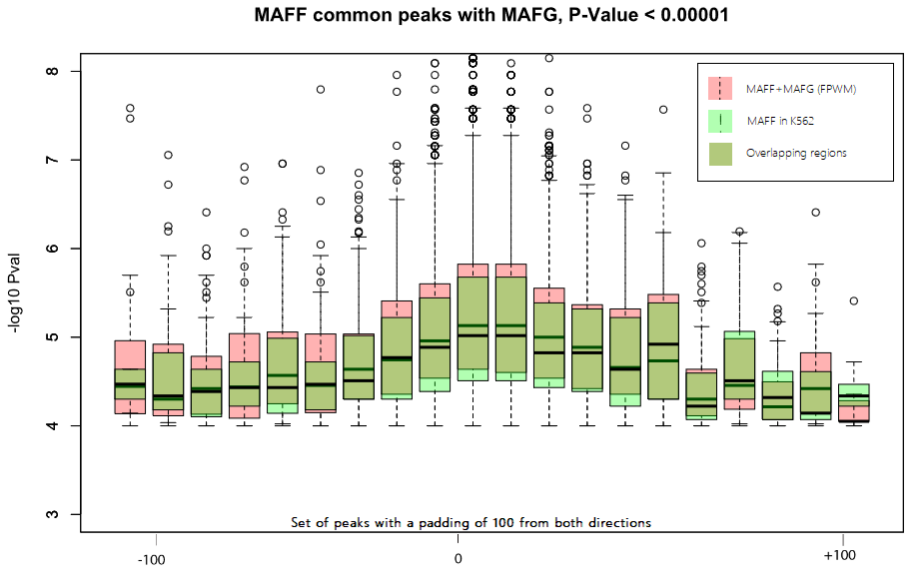


Figure 3.17: FPWM for MAFF-MAFG is used for scanning common peaks of MAFF and MAFG in the K562: In here, we are showcasing another dimer (MAFF-MAFG homodimer matrix from FPWM) in the K562. The performance of two different matrices is compared here. The FPWM of MAFF+MAFG (shown in pink) and the global matrix of MAFF (shown in green) are used to scan the set of peaks that MAFF and MAFG are sharing (which corresponds to the MAFF-MAFG homodimer). As it is depicted, the FPWM shows a better match (higher P-value) as the scanner approaches towards the center of TFBS (o on the x-axis). This shows the better prediction power of FPWM compared to a general matrix for all dimers in a given cell-line.

The same result comes out when we study another dimer in the same cell line which can be seen in next figure.

In 3.18 it is illustrated how the FPWM for MAFF-NFE2 dimer, shows better binding

site prediction compared to global matrix for MAFF, which is the target TF in this case.

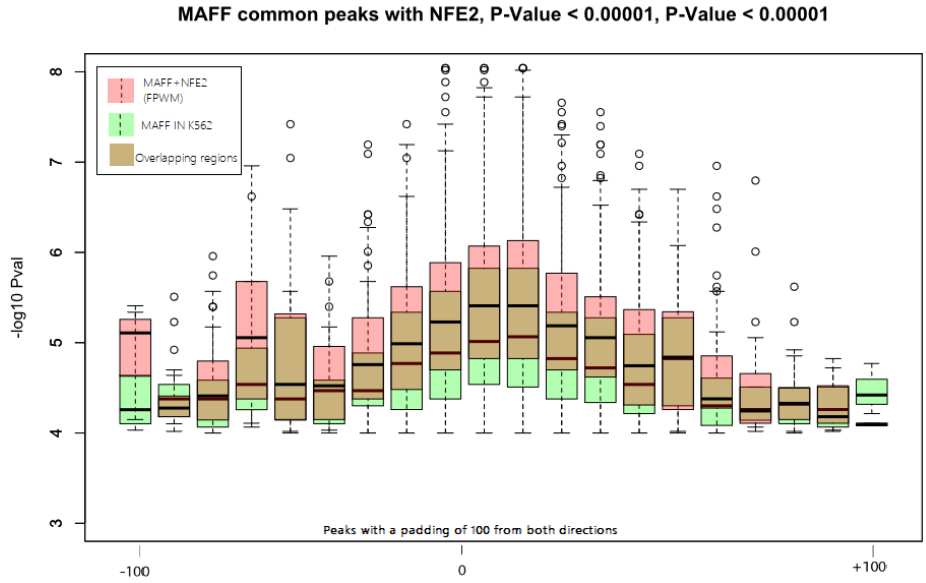


Figure 3.18: MAFF and NFE2 case in K562: Here, we are showcasing another dimer (MAFF-NFE2 heterodimer matrix from FPWM) in the K562 to prove the outperformance of FPWM for both homodimer and heterodimers in a given cell line for all the dimer complexes. The performance of FPWM of MAFF+NFE2 (shown in pink) and the global matrix of MAFF (shown in green) are compared through scanning the set of peaks that MAFF and NFE2 are sharing (corresponding to the MAFF-NFE2 heterodimer). As seen, the FPWM results in a better match (higher P-value) as the scanner approaches towards the center of TFBS (o on the x-axis). This means better prediction power of FPWM than a general matrix for all dimers in a given cell-line.

3.2.6 FPWM of a targeted TF in different cell lines

The FPWM is not only a more powerful tool for different dimers of a TF, but it also outperforms when it comes to the prediction of dimer binding sites in different cell lines. In this subsection, we will explain how studying a target TF in different cell lines can result in different sequence logos. Then it will be shown how a cell-specific FPWM of a given TF can enhance prediction power.

To better portrait, the impact of different cell environments on the sequence preference

of a dimer, an example of CEBPB as a target TF is shown in the following figures. Here, we aim to show how augmenting the motif of a certain TF in a given cell line can deform the visualized motif profile. For example, consider the below figure:

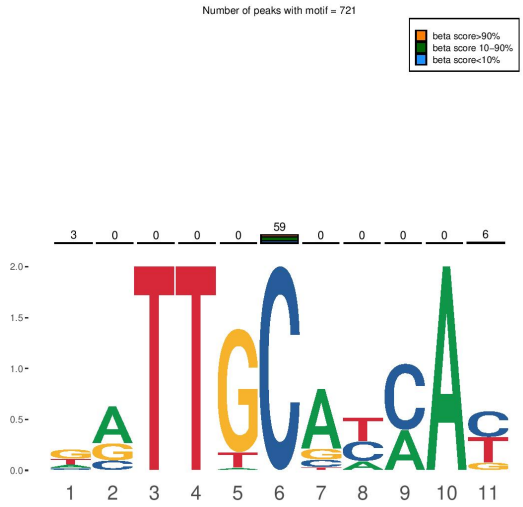


Figure 3.19: CEBPB motif profile in a single cell-line (K562): the sequence logo is coupled with the methylation profile of this TFBS, as can be seen on top of the figure. Methylation profile plays a big role in genome occupancy of a given TF.

The 3.19 is the motif profile for CEBPB in K562 only. However, if another cell line that shares the same peaks are added to the set of sequences to construct the motif profile, we will observe some changes in one part of the sequence logo coupled with methylation level as shown in 3.20:

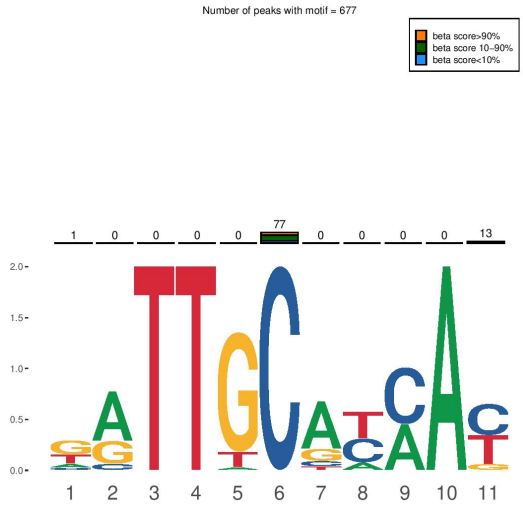


Figure 3.20: CEBPB motif profile in K562, with taking intersected peaks coming from an other cell-line. Note how involvement of only one extra cell-line puts an impact on both DNA and methylation profile. (refer to Figure 3.19 for better understanding). As can be seen, the methylation level is increased in positions 6 and 11, which can alter the sequence preference of the given TF in bigger scales.

As can be seen, adding only one cell line's common peaks with the current one (K562) alters the DNA sequence profile along with the methylation level. In the 3.21, we show the same event, for the case of REST.

As shown in Figure 3.21, a motif profile for this TF shows a different pattern, especially in the case of spacer nucleotides. Furthermore, it is clear that all sequence logos are conserved in the left half, and the second half of them are under the influence of the cell lines differentiation. Regarding this behaviour, a graph of the sequence logos would be a more efficient visualization. In the figure, we represent q forked PWM for the TF named "REST." From the topmost logo towards down, the new cell line peaks are added to the previous one forming a similar but an altered sequence logo. Note that this is different from the intersection matrix because more and more data coming from cell lines are taken into account. After adding all cell lines, the final logo would be the general sequence logo for a given TF, without any cell line specifications.

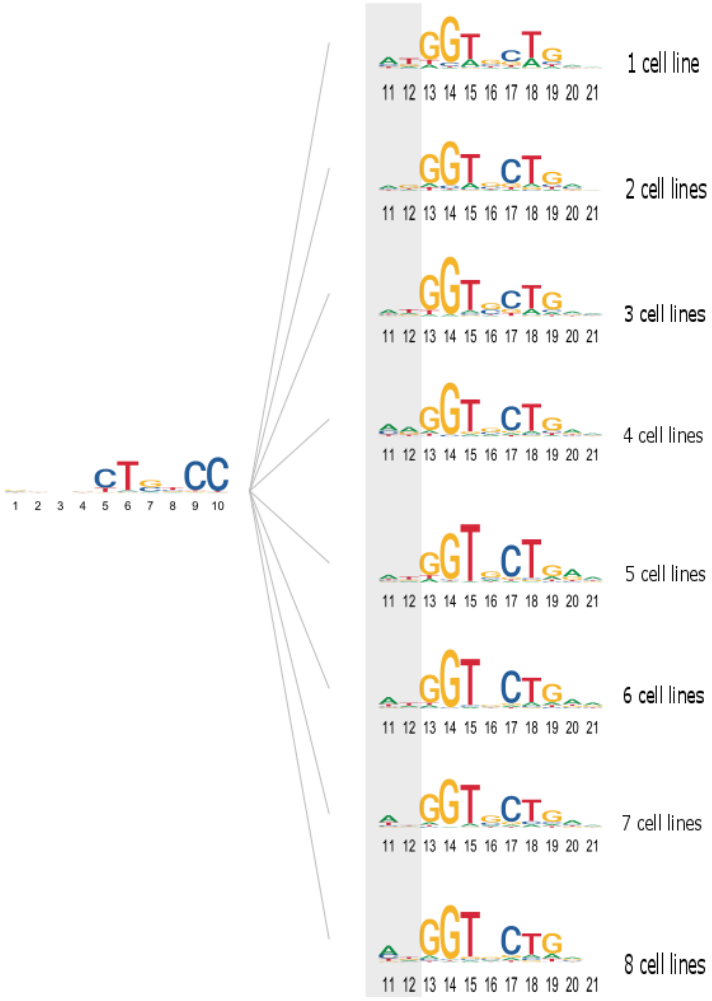


Figure 3.21: A forked sequence logo for REST by augmenting the initial matrix profile (REST in K562) with additional cell lines for the same TF. In this plot, the differentiation of the DNA profile for REST is showcased. The plot on the left side is the conserved region of TFBS that barely changes when a new set of peaks from a new cell-line is added into the current data. However, by taking a closer look at the spacer nucleotides at positions 11-12 (highlighted in grey), it is clear how spacers are toggling with the addition of more cell lines.

As shown in Figure 3.21, as the data coming from a new cell line adds to the previous one, it leads to some alteration, which is more obvious in spacer nucleotides (positions 11 and 12). For example, starting with REST in K562, you can observe that spacer nucleotides are AT. By adding one cell line to this, these nucleotides are changed to AG. By adding two other cell lines, it is observed that the spacers turn

to AA. This happens along with minor modifications to the rest of the sequence logo. This observation motivates us to study the FPWM of a given TF but in different cell lines. For that purpose, we analyze the TF in cell line GM12878, forked to JUND (the previous step is implied) in cells in HCT116, K562, H1-hESC and HepG2. Here we only present the result for FPWM of JUND in cells in GM12878 forked to HCT116.

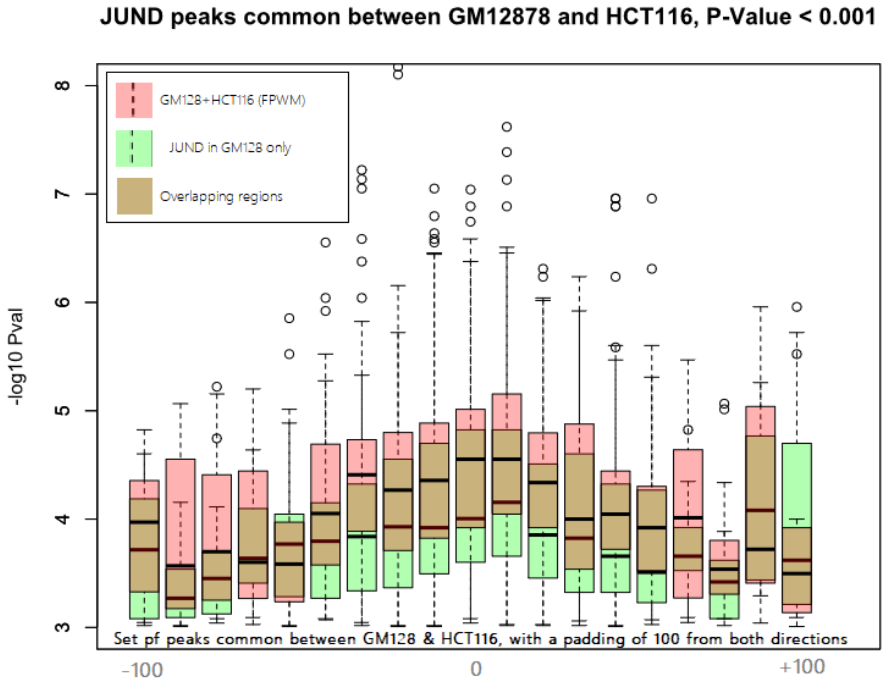


Figure 3.22: The result of scanning JUND peaks shared between cell lines GM12878 and HCT116 using FPWM. Here we represent the performance of an FPWM that is constructed for a given TF but multiple cell-lines. For this purpose, an FPWM of cell-lines GM12878 and HCT116 is constructed for JUND. Then, the set of peaks shared between these two cell-lines are scanned using both FPWM (in pink) and the global motif profile of JUND in GM12878. As expected, the FPWM shows a better match compared to the global matrix.

As it is presented in 3.22, the FPWM, shows better scores compared to the general matrix profile. This visualization proves that we successfully enhanced TFBS prediction resolution, even more, using cell-specific motif profiles. This is considered a second layer of enhancement in prediction power, which works perfectly for dimer

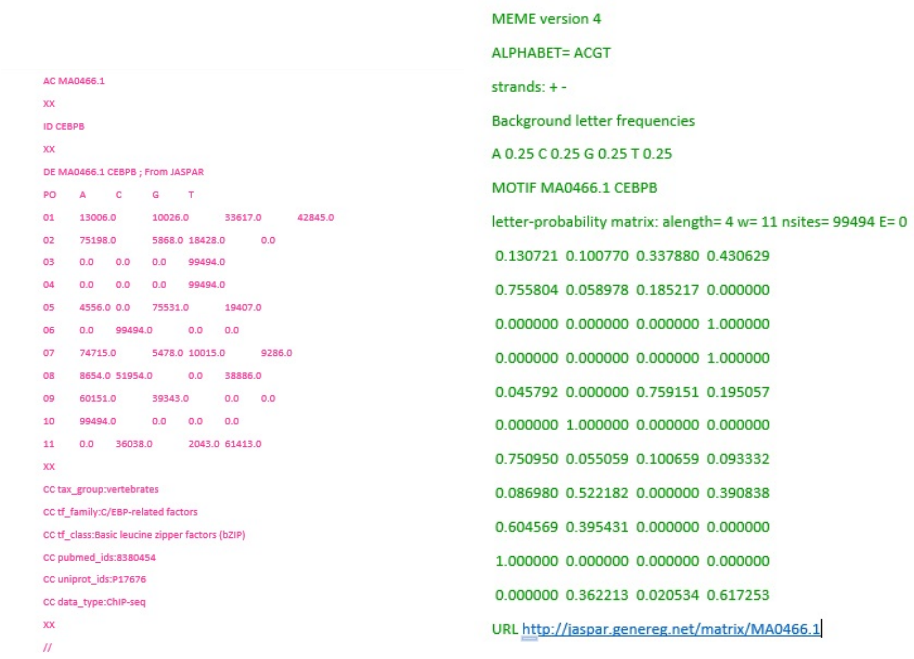
complexes.

In this chapter, we observed the behaviour of the FPWM and how the matrix performs in the context of different co-factors of a TF in a given cell line, or different cell lines in which a given TF exists. Through this, we showed how different partnering, or environmental platform (of each cell line) can alter the sequence preference, and how the FPWM is a better tool for predicting dimer TFBSs. This also imply a more specific and noise-free prediction requires the assessing tools to increase their accuracy and change their perspectives when dealing with different TF families, a problem which FPWM addresses and improves.

3.2.7 Complimentary Analysis

RSAT performance evaluation

As it was explained before, the data file formats are a visual structure for holding specific information. Two popular data formats for motif profile matrices are TRANSFAC and MEME. By convention, TRANSFAC holds a PCM matrix to represent motif profile, while MEME keeps a PPM. The case here is that data format should not be one of the factors that impact TFBS scanning and matrix evaluation. However, throughout our studies, it was observed that RSAT matrix scanning shows some irregularities. Due to the importance of proper data formatting to us, we designed a targeted exam in order to study the behaviour of RSAT towards different data formats. For this purpose, we selected the matrix profile for CEBPB from JASPAR, via the RSAT website, as seen in Figure 3.23.



```
AC MA0466.1
XX
ID CEBPB
XX
DE MA0466.1 CEBPB ; From JASPAR
PO       A       C       G       T
01     13006.0         10026.0         33617.0         42845.0
02     75198.0         5868.0  18428.0         0.0
03     0.0     0.0     0.0     99494.0
04     0.0     0.0     0.0     99494.0
05     4556.0  0.0     75531.0         19407.0
06     0.0     99494.0         0.0     0.0
07     74715.0         5478.0  10015.0         9286.0
08     8654.0  51954.0         0.0     38886.0
09     60151.0         39343.0         0.0     0.0
10     99494.0         0.0     0.0     0.0
11     0.0     36038.0         2043.0  61413.0
XX
CC tax_group:vertebrates
CC tf_family:C/EBP-related factors
CC tf_class:Basic leucine zipper factors (bZIP)
CC pubmed_ids:8380454
CC uniprot_ids:P17676
CC data_type:ChIP-seq
XX
//
```

```
MEME version 4
ALPHABET= ACGT
strands: + -
Background letter frequencies
A 0.25 C 0.25 G 0.25 T 0.25
MOTIF MA0466.1 CEBPB
letter-probability matrix: alength= 4 w= 11 nsites= 99494 E= 0
 0.130721  0.100770  0.337880  0.430629
 0.755804  0.058978  0.185217  0.000000
 0.000000  0.000000  0.000000  1.000000
 0.000000  0.000000  0.000000  1.000000
 0.045792  0.000000  0.759151  0.195057
 0.000000  1.000000  0.000000  0.000000
 0.750950  0.055059  0.100659  0.093332
 0.086980  0.522182  0.000000  0.390838
 0.604569  0.395431  0.000000  0.000000
 1.000000  0.000000  0.000000  0.000000
 0.000000  0.362213  0.020534  0.617253
URL http://jaspar.genereg.net/matrix/MA0466.1
```

Figure 3.23: Motif profile matrices of CEBPB from JASPAR, in TRANSFAC (left) and MEME (right) format. Both data files start with some informative lines, followed by a motif profile matrix. In the TRANSFAC format (on the left), the motif profile matrix is a PCM, and the sum of columns in each row is the same number of studied sequences. In MEME format (on the right), the main matrix is a PPM in which the sum of columns is equal to one.

To determine whether the data format alters the final result, we format the PPM in TRANSFAC format. To be more specific, we replace the motif profile matrix of TRANSFAC data with the matrix from the corresponding MEME file. This is basically a simple normalization of PCM to PPM, as shown in Figure 3.24.

```
AC MA0466.1
XX
ID CEBPB
XX
DE MA0466.1 CEBPB ; From JASPAR
PO       A       C       G       T
01 0.130721 0.100770 0.337880 0.430629
02 0.755804 0.058978 0.185217 0.000000
03 0.000000 0.000000 0.000000 1.000000
04 0.000000 0.000000 0.000000 1.000000
05 0.045792 0.000000 0.759151 0.195057
06 0.000000 1.000000 0.000000 0.000000
07 0.750950 0.055059 0.100659 0.093332
08 0.086980 0.522182 0.000000 0.390838
09 0.604569 0.395431 0.000000 0.000000
10 1.000000 0.000000 0.000000 0.000000
11 0.000000 0.362213 0.020534 0.617253
XX
CC tax_group:vertebrates
CC tf_family:C/EBP-related factors
CC tf_class:Basic leucine zipper factors (bZIP)
CC pubmed_ids:8380454
CC uniprot_ids:P17676
CC data_type:ChIP-seq
XX
//
```

Figure 3.24: A normalized TRANSFAC data format, as the result of merging the standard TRANSFAC and MEME formats presented in 3.23. Here, the general structure of the TRANSFAC format is kept (shown in pink), but the representative matrix is substituted with the corresponding matrix from MEME format (shown in green), which is a PPM (normalized to the scale of 1).

Three of these matrices are representing the same TFBS only in different data formats. It is expected that the result of the evaluation would be the same for all these matrices. However, the result of the evaluation shows otherwise.

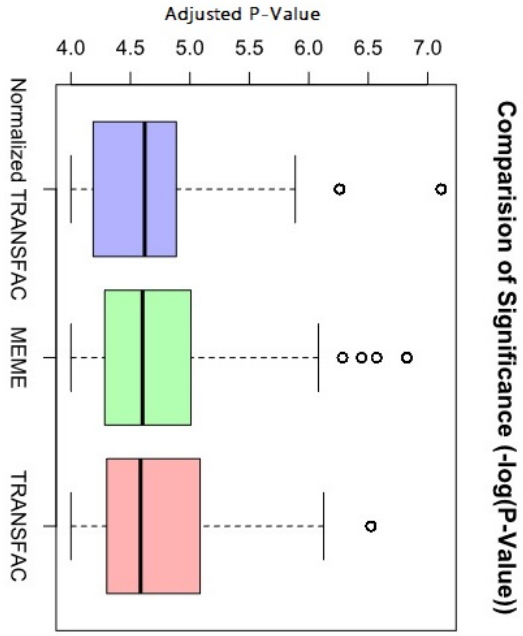


Figure 3.25: Comparison of RSAT's matrix scanning using three matrices for CEBPB binding site motif. The TRANSFAC file is holding a PCM matrix with the sum of rows equal to the number of sequences involved in the study. In MEME format, the main matrix is a PPM with the sum of rows equal to 1. The normalized TRANSFAC format has the same data format as the TRANSFAC file but holds the same representative matrix as MEME format (a PPM). Here, we observed slightly different P-values for each of these files, which shows a data-format-oriented behaviour of RSAT, which is inadequate.

figure 3.25 represents a box plot of the P-values of each matrix when used to scan CEBPB binding sites. As it is shown, although they are quite the same, they are not identical. This issue is even more critical in MEME and normalized TRANSFAC, as they are holding the same matrix. In the following figure, the weight score of the matrices are represented as a box plot.

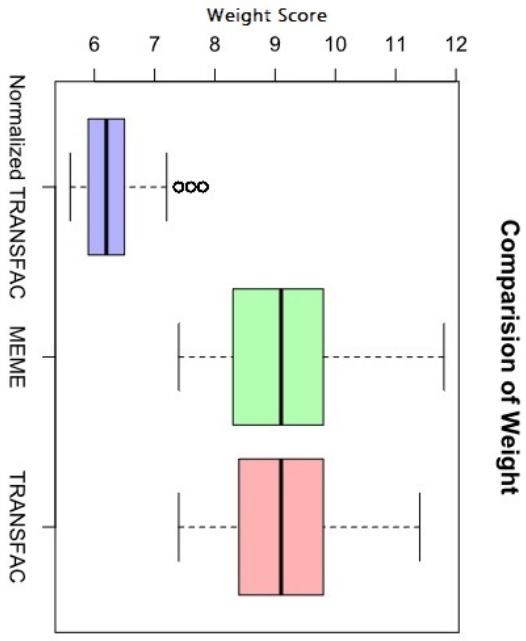


Figure 3.26: Comparison of RSAT's matrix scanning using three matrices for CEBPB binding site motif. As mentioned earlier, The TRANSFAC file is holding a PCM matrix with the sum of rows equal to the number of sequences involved in the study. In MEME format, the main matrix is a PPM with the sum of rows equal to 1. The normalized TRANSFAC format has the same data format as the TRANSFAC file but holds the same representative matrix as MEME format (a PPM). Here, we represent a dramatic drop in the normalized TRANSFAC file's weigh score compared to the other file formats holding basically the same matrix.

As shown in 3.26, the weight score of the matrices also differ one from another. This difference is more significant in the case of normalized TRANSFAC. As explained, although it holds the same matrix that MEME format does, it shows a dramatically lower weight compared to the other two. Regarding all these, we can conclude that RSAT's matrix scanner makes an assumption based on data format and behaves differently regarding what is the matrix format. To avoid this problem in our results, we exported and used all the matrices of the TRANSFAC format holding a PCM within them, then used the P-value as our metric for comparison. However, this issue motivated us to develop our own matrix scanner, which would be our next step of

research.

Data format and compatibility

One of the biggest challenges throughout this study was dealing with names and definitions that are interchangeably employed to refer to varied types of TFBS matrices. It is undeniable that general terms to address equivalent but distinct subjects can impact downstream analysis and final interpretation. It can also put massive overhead on the procedure of linking various components that were not compromised on terms and definitions in the first place. For instance, it is a common -but problematic- practice to use the term PFM to refer to either the Position Count Matrix (PCM) or Position Probability Matrix (PPM). Such imprecise usage of words leads to random selections of data format, causing incompatibility among tools. Moreover, this inconsistency puts the massive overhead of format conversion on the user, which may add a bias to the final result and increase the risk of human error and restricts the user to the compatible tools rather than the outperforming ones. More importantly, this keeps the researcher from comparing tools and applications in a systematically proper way. The input/output data format is typically determined in the application design phase and before developing. Regarding the lack of standard definition for data formats, developers either have to limit themselves to several target tools that they want to be compatible with or cover all the possible data formats, which is computationally expensive. Regarding these issues, we needed to detect the most common formats, evaluate it, and then develop our program so that the output can be widely employed. In previous sections, it was already explained how different matrix profiles could be generated from a set of aligned sequences to represent a collection of targeted sequences systematically. Also, in previous sections, it was already explained how different matrix profiles could be generated from a set of aligned sequences to

represent a collection of targeted sequences. From a FASTA file of aligned sequences, we can directly generate a Position Count Matrix. Each column of this matrix will correspond to each of the positions, while each element is holding each nucleotide's count, associated with that very row. As may be obvious, the sum of columns in PCM is equal to the number of sequences under study. To unify study cases and ease the comparison and interpretation, it would be of the help the whole matrix is normalized in such manners that the effect of the count is discarded from the result. Thus, the PCM's elements can be divided by the total number of the count, resulting in the Position Probability Matrix, which holds elements all smaller than 1. It is needless to say, the sum of the columns in PPM would and should be equal to 1 regardless of the number of sequences under study. Regarding different representations of the motif, profiles have been explained on page 53 and how they end up in sequence logos. However, different data formats contain a matrix profile and some additional data for user knowledge. We studied several tools and applications in this field to make a precise decision on the right approach in this case.

The two major data formats being employed by popular tools and databases are TRANSFAC [84] and MEME [85].

However, due to the lack of outlined protocols for data formats, many applications prefer working with a raw matrix profile instead of formatting it in a conventional structure. Before describing these data formats' anatomy, it would help distinguish between several most used terminologies to better understand how a motif matrix can be represented, then formatted into a proper data file.

To precisely describe the existing issue in the sense of data format, it is essential to compromise the choice of words. Different terminologies are frequently used to refer to distinct formats, and in here, we are trying to avoid making the same mistake by reviewing several terms before falling into the main issue we are going to address.

The first term, which is constantly used in this area, is "Frequency." According to Merriam-Webster, the frequency can be defined as "the number, proportion, or percentage of items in a particular category in a set of data [3]." regarding that and since the term "number," "proportion," and "percentage" refer to a distinct type of matrices, we find this term too general and not quite suitable to be used for referring to matrices. To clarify further, reviewing the terms "count" and "probability" is needed. Merriam-Webster defines the "count" as "a total obtained by counting and the probability as "the ratio of the number of outcomes in an exhaustive set of equally likely outcomes that produce a given event to the total number of possible outcomes [86]." This proportional nature of probability sets a range on it (normally 0 to 1), which is not the case in counting. However, due to the more intuitive image that percentage can give, this number is typically normalized over 100, resulting in "percentage." By aforementioned, two important conclusions can be drawn:

1. The term frequency (frequency matrix in particular) is a general term, and its employment can cause in the random selection of either of those three matrices mentioned above.

2. Although the count matrix and probability matrix represent the same motif profile, the count can be any integer number, while probability has a range of [0,1].

We rather use “percentage matrix” instead of probability matrices, which are normalized over 100, only for the sake of precision. Another thing to be pointed out is PWM usage to refer to all kinds of matrices representing a motif profile. Although this terminology is not wrong, again, to be more precise, it should be noted that PWM is a particular matrix that is directly used to construct a sequence logo. However, PWM is widely used to refer to all types of motif profile matrices. Although we accept this

convention here, if we look at the visual presentation of a probability matrix and a PWM matrix, we can see how these two are indeed different types of matrices.

Following this, we studied several popular tools to investigate what type of data format is more popular among researchers. The summary of this review can be seen in the table 3.1

Output's tool name	PCM	PPM	PWM	Formatted data file
JASPAR	✓			TRANSFAC, MEME
HOMER[87]		✓		
HOCOMOCO	✓		✓	
FACTORBOOK[88]	✓	✓		
RSAT				TRANSFAC
SwissRegulon[89]	✓	✓		
MethMotif*				TRANSFAC, MEME
TFregulomeR*				TRANSFAC, MEME
FootprintDB[90]	✓	✓		

Table 3.1: Quick review of selected applications to investigate their preferred data formats. PCM, PPM and PWM represents pure matrices. As can be observed, many databases prefer to provide the user with only matrices rather than formatting them into common data file formats such as TRANSFAC and MEME. *MethMotif and TFregulomeR() are discussed further in the text.

As it is shown, TRANSFAC and MEME are the most common data formats that have been accepted amount developers in this field. However, the tendency to use raw matrix profiles without structurally formatting them is evidence of the lack of strong conventions over data formatting. In the following lines, TRANSFAC and MEME will be further discussed.

Anatomy of selected data formats

TRANSFAC is a data format first introduced by the TRANSFAC database, one of the most well-known databases of the TFs. TRANSFAC has been around for years and has helped significantly to publish data on eukaryotic gene transcription regulation. TRANSFAC's main database is experimentally evident, and its contributions have made available data compatible with many tools and usable for further analysis[91]. TRANSFAC data format is one of the most stable and widely employed formats in databases that most tools and applications employ. The original data format introduced by the TRANSFAC team is shown in 3.27.

TRANSFAC accession number: M00127
TRANSFAC identifier: VSGATA1_03
Name: GATA-1
Description: GATA-binding factor 1

Position	A	C	G	T	Consensus sequence
1	4	1	2	0	R
2	1	1	3	2	N
3	1	2	4	0	S
4	2	2	2	1	N
5	3	0	2	2	D
6	0	0	12	0	G
7	12	0	0	0	A
8	0	0	0	12	T
9	12	0	0	0	A
10	8	1	3	0	A
11	1	4	4	3	N
12	3	4	3	2	N
13	3	1	7	1	G
14	2	4	4	2	N

Statistical basis: 12 selected binding sequences

Figure 3.27: First published data format by TRANSFAC team, taken from [92], openly accessible for public on October 2020.

As shown in the figure, the TRANSFAC format starts with several informative lines and ends with them. The main body that holds the matrix is a PCM matrix with defined columns for each of four types of nucleotides. In addition to those, a column for positions is observed, which represents to which position in the aligned set of sentences the number corresponds. Another column named "Consensus sequence,"

which aims to show the optimum binding site by representing either the most frequent nucleotide at that position or an approximation of the frequency at that very position. The other data format to be discussed is MEME format. This format is used by the MEME suite, an online toolkit that provides users with functions to study and analyze representative sequences such as DNA or protein sequences. The MEME data format is shown in Figure 3.28

```
MEME version 4

ALPHABET= ACGT

strands: + -

Background letter frequencies
A 0.303 C 0.183 G 0.209 T 0.306

MOTIF crp
letter-probability matrix: alength= 4 w= 19 nsites= 17 E= 4.1e-009
 0.000000  0.176471  0.000000  0.823529
 0.000000  0.058824  0.647059  0.294118
 0.000000  0.058824  0.000000  0.941176
 0.176471  0.000000  0.764706  0.058824
 0.823529  0.058824  0.000000  0.117647
 0.294118  0.176471  0.176471  0.352941
 0.294118  0.352941  0.235294  0.117647
 0.117647  0.235294  0.352941  0.294118
 0.529412  0.000000  0.176471  0.294118
 0.058824  0.235294  0.588235  0.117647
 0.176471  0.235294  0.294118  0.294118
 0.000000  0.058824  0.117647  0.823529
 0.058824  0.882353  0.000000  0.058824
 0.764706  0.000000  0.176471  0.058824
 0.058824  0.882353  0.000000  0.058824
 0.823529  0.058824  0.058824  0.058824
 0.176471  0.411765  0.058824  0.352941
 0.411765  0.000000  0.000000  0.588235
 0.352941  0.058824  0.000000  0.588235
```

Figure 3.28: An example of MEME data format, taken from Minimal MEME format examples

As can be seen, this data format also starts with several descriptive columns. Then comes the main body of data, which holds the motif profile. This format, by convention, holds a PPM without any additional information about rows or columns. However, in descriptive lines, it is mentioned in what order the columns are associated with each nucleotide. Also, since the range of numbers is 0 to 1, the number of studied sequences should be named.

With regards to these two data formats, the differences can be narrowed down to the following main points:

1. TRANSFAC format holds a PCM, while MEME format holds a probability matrix.

2. The sum of columns in TRANSFAC format is equal to the number of sequences that have been studied, while in MEME format, it is always equal to 1.

3. MEME format, despite TRANSFAC, does not have an extra row and a column for names and positions, respectively.

Regarding all the aforementioned, it should be clear by now how data formats can vary in details thus resulting in incompatibility among analysing tools and applications. We aim to highlight the importance of creating standard common dialogue among researchers and avoid any potential trouble that can easily result in big impact on the final outcome.

Chapter 4

Discussion and Conclusions

4.1 Discussion

PWMs, along with Sequence Logos, have been the standard tools for modelling and visualizing the TFBSs. These tools rely on a collection of DNA sequences that are targeted by a specific TF. These sequences were originally identified through in vitro experiments. These included a set of gel-shift followed by SELEX, which delivered a set of sequences targeted by a given TF. By progress that high-throughput technologies have brought into this field, it has been possible to study the TFBSs at a genome-wide level in the in vivo context by Chromatin Immunoprecipitation of a given TF, followed by high-throughput sequencing of the bound DNA loci. ChIP-Seq assay allows the researcher to characterize TFBSs of a given TF in the cell context, considering transcription co-factors and chromatin state. These in vitro assays were executed in a regulated condition where only one TF of interest is studied versus a library of sequences. However, there are classes of TF, with a specific structure that binds to the DNA sequence as dimers (e.g. Leucine Zippers). This family of TFs work in groups to target DNA sequences. They partner with a protein similar to them-

selves (resulting in a homodimer) or distinct from them (composing a heterodimer) and bind to DNA sequence as a complex. Thus, the ChIP-Seq assay will capture the entire collection of dimer complexes that are in the cell. With this nature of ChIP-Seq and subsequently aggregating heterogeneous DNA binding sequences in order to form a single PWM to represent TFBSs, results in a systematically noisy matrix or dyed motif that includes a conserved region, which corresponds to the binding sites of TF of interest, and a degenerated part, corresponding to the aggregated binding sites for partners of TF of interest. Regarding the abovementioned, it can be concluded that traditional representative models such as PWM and sequence logos do not properly represent the TFBSs and do not accurately model binding sites for dimer complexes, as the degenerated part of representation is noisy. To tackle this issue, we have developed FPWM, an R-Library that provides users with functionalities for generating Forked PWMs and Forked sequence logos. This library enables the users to visualize a more expressing plot that reflects the background scenario of dimers and their partners, along with matrices representing the sequence affinity of a given TF with those of a segregated list of partners more accurately. FPWM, contrary to current modelling methods, represents dimer partners as PWMs and Sequence logos, forked from the main TF motif. The current version of FPWM uses the TFregulomeR() library to identify co-factors and to deconvolute a given TF's partners. FPWM explores and delivers a report of the given TF's partners in a cell-specific approach, ranked based on their amount of co-binding, along with their Sequence Logos. This report helps the user study and analyze the significant partners and identify a proper forking point for downstream analysis. After determining the forking position (in the sense of DNA sequence and/or methylation profile) and exporting the list of top co-binding partners, the user can generate a forked PWM (FPWM) then a forked sequence logo corresponding to the matrix. The FPWM can be used for matrix-based analysis, and

the plot takes a more expressive approach to imply the dimer nature of the given TF, along with reflecting the information about the binding partner and the percentage of the co-binding level. In addition to the TF partners' study, FPWM can generate matrices for a given TF, forked into the cell lines currently in the database. This process adds another layer of improvement into the cell-line specific TFBS matrices and enhances TFBS prediction power in several different cell lines that the TF has been observed in. Analogous to the TF partner studies, a forked model for one specific TF can be generated using the provided functionalities to carefully observe the impact of aggregation of multiple cell lines into one representative matrix, then scanning novel sites with a selective manner.

As mentioned in previous sections, due to the popularity of the TRANSFAC format among the developers of relevant tools and applications, users of the FPWM package are provided with functionalities to export and generate FPWM standard TRANSFAC format. Each data file contains an FPWM along with information about forking position and overlapping percentage for further examinations.

4.2 Conclusions

Using FPWMs for TFBSs prediction on DNA sequences using Matrix Scan tools on the RSAT website results in a more accurate and higher match-score than traditional PWMs representing a given dimer complex. Certainly, traditional PWMs carry more noise and less informative bases in dimer co-binding regions. The FPWM allows the user to scan chimeric motifs composed of a merge of several binding sites coming from different TFs. Since these binding sites are not exhibited in the collection of DNA sequences employed to construct the original PWM, using these matrices can easily end in false-positive predictions. In addition to this, its dimer partner can alter the

function of a specific TF. FPWM represents a novel approach to better model and visualization of the context-specific TFBSs.

To evaluate the Forked PWMs and compare it to its resource database (Meth-Motif's), a set of motif scanning was performed using matrices from FPWM and MethMotif. For this purpose, several TFs from the bZIP family were chosen as TF of interest, followed by overlapping analysis to retrieve the set of sequences common between a TF and its particular partner. The coordinates of peaks co-bound by TF of interest and its partner were loaded as a BED file from MethMotif. The peaks were expanded from both directions with padding of -+100, and the result FASTA file was achieved. This FATSA file was then scanned using matrices from MethMotif and FPWM. The set of matched sequences with the P-Value threshold <= 0.0001 were explored to target the sequence holding the best p-value corresponding to the center of TFBS. As a result, it was observed that FPWM improves PWM models of TF dimers, thus enhancing our understanding of the TF cooperatively and opening a new window to accurate TFBS prediction. In the end, the impact of FPWM, can be summarized as follow:

1. The FPWM is a novel approach to visualize TFBSs more expressively. By using FPWM, the susceptibility to false positives and systematic noises are reduced, and the power of TFBS prediction is enhanced.

2. The FPWM provides a better understanding of a given TFs functionality by allowing the user to have a more precise interpretation of a given TFs binding dynamic.

3. Aside from the computational value of the FPWM, it opens the door for precision in biological aspects by putting the given TF in different contexts: cell-line wise or co-factor wise.

4. The FPWM is embedded into MethMotif database as a new tool for TFBS analysis and is aiming to represent a new standard in exploring TFBSs for dimer families.

4.3 Future extensions and publications

In this project, we enhanced the power of TFBS prediction for the bZIP family, a large family of TFs with core functionalities throughout cell evolution. Malfunction of this group of co-factors is associated with several challenging diseases such as cancer[93] or Leukemia[94]. Regarding the significance of this family and the drawback of current TFBS databases, we hope that current databases address the issue highlighted by us and update their representative models using the FPWM library. We started this update with our lab's database (MethMotif), and we are going to release the next version of MethMotif (MethMotif 2021) with FPWM for dimer TFBSs.

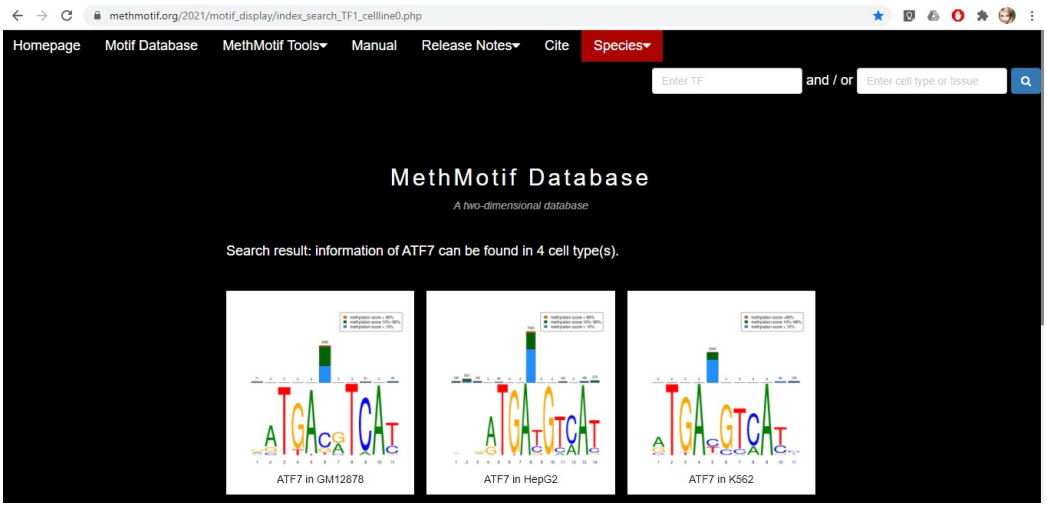


Figure 4.1: A look up for ATF7 in MethMotif 2021 accessible at `methmotif.org/2021`.

Figure 4.2: In details of each profile, FPWM is available for download.

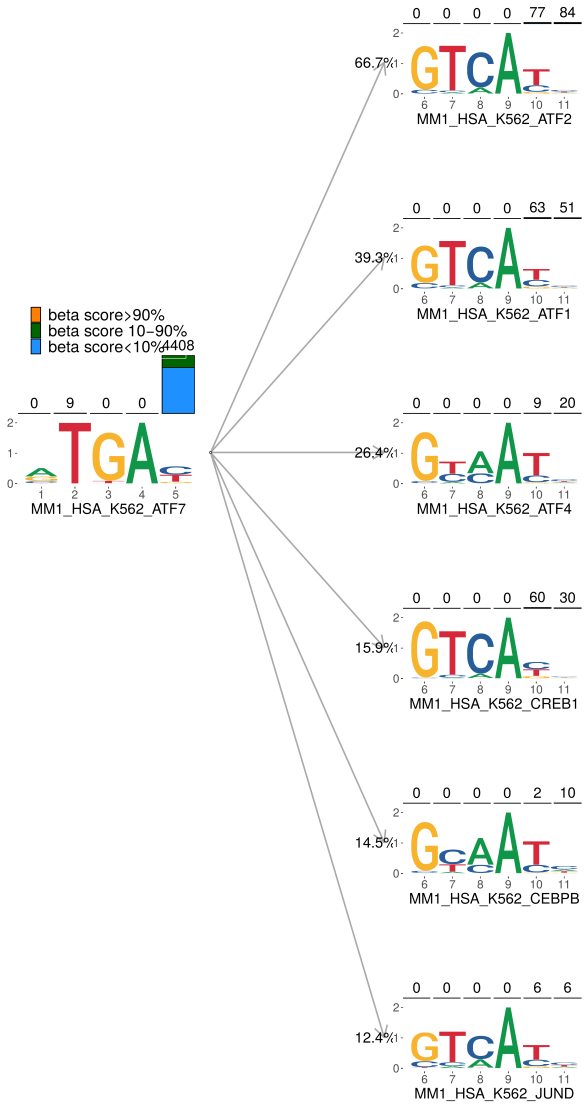


Figure 4.3: The downloaded file shows the plot for number of most significant partners of ATF7.

Hopefully, FPWM will be improved and updated to cover more families of TF in the future. We are also looking forward to dragging the attention of other popular databases to this problem and seeing the update on their websites. Furthermore, due to some issues faced while dealing with matrix scanners, we believe that the current version of matrix scanners highly relies on a simplified structure from TF and do not

perform well enough for more complicated complexes such as TF dimer. When writing this thesis, we are working closely with the team behind RSAT to develop a scanner that is suitable for dimer analysis with higher compatibility with FPWM. We are looking forward to seeing these updates in the next releases of RSAT, and eventually, all the TFBS analysis tools and applications. For dragging public attention, we have published this work in several formats so far, as follows:

1. The manuscript was initially submitted to the journal of Bioinformatics (with an impact factor of 5.61), and currently, we are working on the revisions. The name of contributors to the paper is mentioned in the "Statement of Co-Authorship)" section, outlining the roles of each contributor.

2. This work has been represented as a subdivision of MethMotif in two conferences, namely "Applied Bioinformatics in Life Sciences (3rd edition)", held in Leuven, Belgium, on February 13-14, 2020, and 28th conference on "Intelligent Systems for Molecular Biology (ISMB)," on July 16-13, 2020. The latter one was located in Montreal and held virtually due to known circumstances.

Figure 4.4: Our work has been selected for a poster presentation during the VIB Conference Applied Bioinformatics in Life Sciences (3rd edition).

Figure 4.5: The FPWM was presented at the visual conference on Intelligent Systems for Molecular Biology (ISMB), 28th.

Bibliography

[1] J. D. Watson, F. Crick *et al.*, "A structure for deoxyribose nucleic acid," 1953.

[2] A. Travers and G. Muskhelishvili, "Dna structure and function," *FEBS Journal*, vol. 282, no. 12, pp. 2279–2295, 6 2015. [Online]. Available: https://doi.org/10.1111/febs.13307

[3] "Gene," accessed on September 04, 2019. [Online]. Available: https://www.merriam-webster.com/dictionary/gene

[4] "A scientific illustration of how epigenetic mechanisms can affect health."

[5] G.-W. Li and X. S. Xie, "Central dogma at the single-molecule level in living cells," *Nature*, vol. 475, no. 7356, p. 308–315, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21776076

[6] J. S. Dorman, M. J. Schmella, and S. W. Wesmiller, "Primer in genetics and genomics, article 1: Dna, genes, and chromosomes," Nov 2016, accessed on September 04, 2019. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1099800416678321

[7] P. Heyn, A. T. Kalinka, P. Tomancak, and K. M. Neugebauer, "Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences," *Bioessays*, vol. 37, no. 2, pp. 148–154, 2015.

[8] H. Chen, H. Li, F. Liu, X. Zheng, S. Wang, X. Bo, and W. Shu, "An integrative analysis of tfbs-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape," *Scientific reports*, vol. 5, p. 8465, 2015.

[9] D. A. Kleinjan and V. van Heyningen, "Long-range control of gene expression: emerging mechanisms and disruption in disease," *The American Journal of Human Genetics*, vol. 76, no. 1, pp. 8–32, 2005.

[10] F. Spitz and E. E. Furlong, "Transcription factors: from enhancer binding to developmental control," *Nature reviews genetics*, vol. 13, no. 9, p. 613, 2012.

[11] "Homo sapiens comprehensive model collection." [Online]. Available: http://hocomoco11.autosome.ru/

[12] T. Benoukraf, "Analyse bioinformatique des mécanismes de régulation durant le développement précoce des cellules t," Ph.D. dissertation, 2010, thèse de doctorat dirigée par Ferrier, Pierre Bioinformatique, biochimie structurale et génomique Aix-Marseille 2 2010. [Online]. Available: http://www.theses.fr/2010AIX22043

[13] B. S. Gloss and M. E. Dinger, "Realizing the significance of noncoding functionality in clinical genomics," *Experimental & molecular medicine*, vol. 50, no. 8, pp. 1–8, 2018.

[14] A. F. Palazzo and T. R. Gregory, "The case for junk dna," *PLoS genetics*, vol. 10, no. 5, p. e1004351, 2014.

[15] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan *et al.*, "The encyclopedia of dna elements (encode): data portal update," *Nucleic acids research*, vol. 46, no. D1, pp. D794–D801, 2017.

[16] E. P. Consortium *et al.*, "A user's guide to the encyclopedia of dna elements (encode)," *PLoS biology*, vol. 9, no. 4, p. e1001046, 2011.

[17] M. J. Pazin, "Using the encode resource for functional annotation of genetic variants," *Cold Spring Harbor Protocols*, vol. 2015, no. 6, pp. pdb–top084 988, 2015.

[18] D. C. Rio, "Reverse transcription–polymerase chain reaction," *Cold Spring Harbor Protocols*, vol. 2014, no. 11, pp. pdb–prot080 887, 2014.

[19] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, p. 57, 2009.

[20] J. D. Keene, J. M. Komisarow, and M. B. Friedersdorf, "Rip-chip: the isolation and identification of mrnas, micrornas and protein components of ribonucleoprotein complexes from cell extracts," *Nature protocols*, vol. 1, no. 1, p. 302, 2006.

[21] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell, "Argonaute hits-clip decodes microrna–mrna interaction maps," *Nature*, vol. 460, no. 7254, p. 479, 2009.

[22] Y. Liu, L. Fu, K. Kaufmann, D. Chen, and M. Chen, "A practical guide for dnase-seq data analysis: from data management to common applications," *Briefings in bioinformatics*, 2018.

[23] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, "Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin," *Genome research*, vol. 17, no. 6, pp. 877–885, 2007.

[24] O. Paun, K. J. Verhoeven, and C. L. Richards, "Opportunities and limitations of reduced representation bisulfite sequencing in plant ecological epigenomics," *New Phytologist*, vol. 221, no. 2, pp. 738–742, 2019.

[25] M. J. Fullwood and Y. Ruan, "Chip-based methods for the identification of long-range chromatin interactions," *Journal of cellular biochemistry*, vol. 107, no. 1, pp. 30–39, 2009.

[26] M. Ciechomska, L. Roszkowski, and W. Maslinski, "Dna methylation as a future therapeutic and diagnostic target in rheumatoid arthritis," *Cells*, vol. 8, no. 9, p. 953, 2019.

[27] Q. X. Xuan Lin, S. Sian, O. An, D. Thieffry, S. Jha, and T. Benoukraf, "Methmotif: an integrative cell specific database of transcription factor binding motifs coupled with dna methylation profiles," *Nucleic acids research*, vol. 47, no. D1, pp. D145–D154, 2018.

[28] N. Olova, F. Krueger, S. Andrews, D. Oxley, R. V. Berrens, M. R. Branco, and W. Reik, "Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting dna methylation data," *Genome biology*, vol. 19, no. 1, p. 33, 2018.

[29] M. F. Carey, C. L. Peterson, and S. T. Smale, "Chromatin immunoprecipitation (chip)," *Cold Spring Harbor Protocols*, vol. 2009, no. 9, pp. pdb–prot5279, 2009.

[30] S. Q. Ye, *Big data analysis for bioinformatics and biomedical discoveries*. CRC Press, 2016.

[31] P. Gade and D. V. Kalvakolanu, "Chromatin immunoprecipitation assay as a tool for analyzing transcription factor activity," in *Transcriptional Regulation*. Springer, 2012, pp. 85–104.

[32] P. J. Park, "Chip–seq: advantages and challenges of a maturing technology," *Nature reviews genetics*, vol. 10, no. 10, p. 669, 2009.

[33] E. T. Liu, S. Pott, and M. Huss, "Q&a: Chip-seq technologies and the study of gene regulation," *BMC biology*, vol. 8, no. 1, p. 56, 2010.

[34] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom, "Chip-seq: using high-throughput sequencing to discover protein–dna interactions," *Methods*, vol. 48, no. 3, pp. 240–248, 2009.

[35] B. Langmead, "Aligning short sequencing reads with bowtie," *Current protocols in bioinformatics*, vol. 32, no. 1, pp. 11–7, 2010.

[36] "Galaxy training: Mapping," accessed on September 10, 2019. [Online]. Available: https://shiltemann.github.io/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

[37] R. Thomas, S. Thomas, A. K. Holloway, and K. S. Pollard, "Features that define the best chip-seq peak calling algorithms," *Briefings in bioinformatics*, vol. 18, no. 3, pp. 441–450, 2016.

[38] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li *et al.*, "Model-based analysis of chip-seq (macs)," *Genome biology*, vol. 9, no. 9, pp. 1–9, 2008.

[39] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, "A clustering approach for identification of enriched domains from histone modification chip-seq data," *Bioinformatics*, vol. 25, no. 15, pp. 1952–1958, 2009.

[40] Y. Chen, N. Negre, Q. Li, J. O. Mieczkowska, M. Slattery, T. Liu, Y. Zhang,

T.-K. Kim, H. H. He, J. Zieba *et al.*, "Systematic evaluation of factors influencing chip-seq fidelity," *Nature methods*, vol. 9, no. 6, pp. 609–614, 2012.

[41] D. S. Johnson, W. Li, D. B. Gordon, A. Bhattacharjee, B. Curry, J. Ghosh, L. Brizuela, J. S. Carroll, M. Brown, P. Flicek *et al.*, "Systematic evaluation of variability in chip-chip experiments using predefined dna targets," *Genome research*, vol. 18, no. 3, pp. 393–403, 2008.

[42] C. Schubert, "Technology feature: Chip off the old block: Beyond chromatin immunoprecipitation," Dec 2018. [Online]. Available: https://science.sciencemag.org/content/362/6419/1193.2

[43] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. ii. error probabilities," *Genome research*, vol. 8, no. 3, pp. 186–194, 1998.

[44] I. T. Notes, "Quality scores for next-generation sequencing." [Online]. Available: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

[45] P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants," *Nucleic acids research*, vol. 38, no. 6, pp. 1767–1771, 2009.

[46] "Fastq format," Jul 2019, accessed on September 10, 2019. [Online]. Available: https://en.wikipedia.org/wiki/FASTQ_format

[47] M. Scholz, "Sam file format." [Online]. Available: http://www.metagenomics.wiki/tools/samtools/bam-sam-file-format

[48] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth,

G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[49] E. G. Wilbanks and M. T. Facciotti, "Evaluation of algorithm performance in chip-seq peak detection," *PloS one*, vol. 5, no. 7, p. e11471, 2010.

[50] J. Duan, "Computational analysis of chip-seq data," 2010. [Online]. Available: https://pdfs.semanticscholar.org/0984/595dc3a59f42902e54f1e72158d9d757c9e1.pdf?_ga=2.177330349.921017684.1569599647-282187698.1567531595

[51] P. D'haeseleer, "What are dna sequence motifs?" *Nature biotechnology*, vol. 24, no. 4, p. 423, 2006.

[52] M. Geertz and S. J. Maerkl, "Experimental strategies for studying transcription factor–dna binding specificities," *Briefings in functional genomics*, vol. 9, no. 5-6, pp. 362–373, 2010.

[53] M. K. Das and H.-K. Dai, "A survey of dna motif finding algorithms," in *BMC bioinformatics*, vol. 8, no. 7. BioMed Central, 2007, p. S21.

[54] A. M. Khamis, O. Motwalli, R. Oliva, B. R. Jankovic, Y. A. Medvedeva, H. Ashoor, M. Essack, X. Gao, and V. B. Bajic, "A novel method for improved accuracy of transcription factor binding site prediction," *Nucleic acids research*, vol. 46, no. 12, pp. e72–e72, 2018.

[55] S. Hannenhalli, "Eukaryotic transcription factor binding sites—modeling and integrative search methods," *Bioinformatics*, vol. 24, no. 11, pp. 1325–1331, 2008.

[56] M. L. Bulyk, "Computational prediction of transcription-factor binding site locations," *Genome biology*, vol. 5, no. 1, p. 201, 2003.

[57] X. Xia, "Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction," *Scientifica*, vol. 2012, 2012.

[58] "Position weight matrix," Aug 2019, accessed on September 13, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Position_weight_matrix

[59] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, no. 4, p. 276, 2004.

[60] C. T. Workman, Y. Yin, D. L. Corcoran, T. Ideker, G. D. Stormo, and P. V. Benos, "enoLOGOS: a versatile web tool for energy normalized sequence logos," *Nucleic Acids Research*, vol. 33, no. $\text{suppl}_2, pp.W389 - -W392, 072005$.

[61] T. D. Schneider, "Consensus sequence zen," *Applied bioinformatics*, vol. 1, no. 3, p. 111, 2002.

[62] Z. Gao, L. Liu, and J. Ruan, "Logo2pwm: a tool to convert sequence logo to position weight matrix," *BMC genomics*, vol. 18, no. 6, p. 709, 2017.

[63] R. Sakai and J. Aerts, "Sequence diversity diagram for comparative analysis of multiple sequence alignments," in *BMC proceedings*, vol. 8, no. 2. BioMed Central, 2014, p. S9.

[64] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic acids research*, vol. 18, no. 20, pp. 6097–6100, 1990.

[65] A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, J. Cheneby, S. R. Kulkarni, G. Tan *et al.*, "Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework," *Nucleic acids research*, vol. 46, no. D1, pp. D260–D266, 2017.

[66] P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho *et al.*, "The ucsc genome browser database: update 2011," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D876–D882, 2010.

[67] I. V. Kulakovskiy, Y. A. Medvedeva, U. Schaefer, A. S. Kasianov, I. E. Vorontsov, V. B. Bajic, and V. J. Makeev, "Hocomoco: a comprehensive collection of human transcription factor binding sites models," *Nucleic acids research*, vol. 41, no. D1, pp. D195–D202, 2012.

[68] G. Tan and B. Lenhard, "Tfbstools: an r/bioconductor package for transcription factor binding site analysis," *Bioinformatics*, vol. 32, no. 10, pp. 1555–1556, 2016.

[69] A. Medina-Rivera, M. Defrance, O. Sand, C. Herrmann, J. A. Castro-Mondragon, J. Delerce, S. Jaeger, C. Blanchet, P. Vincens, C. Caron *et al.*, "Rsat 2015: regulatory sequence analysis tools," *Nucleic acids research*, vol. 43, no. W1, pp. W50–W56, 2015.

[70] M. Esteller, "Epigenetics in cancer," *New England Journal of Medicine*, vol. 358, no. 11, pp. 1148–1159, 2008.

[71] C. H. Waddington, "Canalization of development and the inheritance of acquired characters," *Nature*, vol. 150, no. 3811, p. 563, 1942.

[72] "Cpg site," Sep 2019, accessed on September 13, 2019. [Online]. Available: https://en.wikipedia.org/wiki/CpG_site

[73] E. Clough and T. Barrett, "The gene expression omnibus database," in *Statistical Genomics*. Springer, 2016, pp. 93–110.

[74] "An integrative cell-specific database of transcription factor binding motifs coupled

with dna methylation profiles," accessed on September 15, 2019. [Online]. Available: http://bioinfo-csi.nus.edu.sg/methmotif/

[75] I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov, "Gtrd: a database on gene transcription regulation—2019 update," *Nucleic acids research*, vol. 47, no. D1, pp. D100–D105, 2018.

[76] L. Chen, C. Wang, Z. S. Qin, and H. Wu, "A novel statistical method for quantitative comparison of multiple chip-seq datasets," *Bioinformatics*, vol. 31, no. 12, pp. 1889–1896, 2015.

[77] S. Inukai, K. H. Kock, and M. L. Bulyk, "Transcription factor–dna binding: beyond binding site motifs," *Current opinion in genetics & development*, vol. 43, pp. 110–119, 2017.

[78] W. Hu, L. Wang, W. Tie, Y. Yan, Z. Ding, J. Liu, M. Li, M. Peng, B. Xu, and Z. Jin, "Genome-wide analyses of the bzip family reveal their involvement in the development, ripening and abiotic stress response in banana," *Scientific reports*, vol. 6, p. 30203, 2016.

[79] J. A. Rodriguez-Martinez, A. W. Reinke, D. Bhimsaria, A. E. Keating, and A. Z. Ansari, "Combinatorial bzip dimers display complex dna-binding specificity landscapes," *Elife*, vol. 6, p. e19272, 2017.

[80] C. Marco Llorca, M. Potschin, and U. Zentgraf, "bzips and wrkys: two large transcription factor families executing two different functional strategies," *Frontiers in plant science*, vol. 5, p. 169, 04 2014.

[81] Q. Lin, D. Thieffry, S. Jha, and T. Benoukraf, "TFregulomeR reveals transcription factors' context-specific features and functions," *Nucleic Acids Research*, vol. 48, no. 2, pp. e10–e10, 11 2019. [Online]. Available: https://doi.org/10.1093/nar/gkz1088

[82] "Frequently asked questions: Data file formats." [Online]. Available: https://genome.ucsc.edu/FAQ/FAQformat.html#format1

[83] I. Albert, "Genomic intervals," 2014. [Online]. Available: https://angus.readthedocs.io/en/2014/_static/2014-lecture4-genomic-intervals.pdf

[84] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, "Transfac: a database on transcription factors and their dna binding sites," *Nucleic acids research*, vol. 24, no. 1, pp. 238–241, 1996.

[85] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The meme suite," *Nucleic acids research*, vol. 43, no. W1, pp. W39–W49, 2015.

[86] "Frequency." [Online]. Available: https://www.merriam-webster.com/dictionary/count

[87] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities," *Molecular cell*, vol. 38, no. 4, pp. 576–589, 2010.

[88] J. Wang, J. Zhuang, S. Iyer, X.-Y. Lin, M. C. Greven, B.-H. Kim, J. Moore, B. G. Pierce, X. Dong, D. Virgil *et al.*, "Factorbook. org: a wiki-based database for transcription factor-binding data generated by the encode consortium," *Nucleic acids research*, vol. 41, no. D1, pp. D171–D176, 2012.

[89] M. Pachkov, I. Erb, N. Molina, and E. Van Nimwegen, "Swissregulon: a database of genome-wide annotations of regulatory sites," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D127–D131, 2007.

[90] A. Sebastian and B. Contreras-Moreira, "footprintdb: a database of transcription factors with annotated cis elements and binding interfaces," *Bioinformatics*, vol. 30, no. 2, pp. 258–265, 2014.

[91] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer *et al.*, "Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D108–D110, 2006.

[92] Z. Zhang and M. Gerstein, "Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements," *Journal of Biology*, vol. 2, no. 2, pp. 1–4, 2003.

[93] T. Ryu, J. Jung, S. Lee, H. J. Nam, S. W. Hong, J. W. Yoo, D.-k. Lee, and D. Lee, "bzipdb: a database of regulatory information for human bzip transcription factors," *BMC genomics*, vol. 8, no. 1, p. 136, 2007.

[94] I. Paz-Priel and A. Friedman, "C/ebpα dysregulation in aml and all," *Critical Reviews™ in Oncogenesis*, vol. 16, no. 1-2, 2011.

Appendix A

Documentation of the FPWM Package

The following is the manual file of the FPWM package, which is a short instruction guideline to work with functionalities. The vignette file of FPWM, along with the recent version of the script is available online at `https://github.com/aidaghayour/FPWM`.

Package

October 15, 2020

Title Forked Position Weight Matrix.

Version 0.0.0.9000

Description This package generates a Froked Position Weight Matrix which is helpful to have a better insight about characteristics of Transcription Factor Dimers.

Depends R (>= 3.5.2)

License What license is it under?

Encoding UTF-8

LazyData true

Suggests knitr, rmarkdown

VignetteBuilder knitr

Imports ggplotify,
ggplot2,
gridExtra,
grid,
lattice,
gridGraphics,
base2grob,
ggplot2,
ggseqlogo,
stringr,
cowplot,
reader

RoxygenNote 7.1.0

R topics documented:

`Barandseqlogo` *A function for generating barchart and seq-logo of co-factors of selected TF*

Description

This function generates a barchart of co-binding Percentage for each co-factor of selected TF, along with seq-logo for each of co-factors.

Usage

```
Barandseqlogo(
  NumberofTop,
  highestscore,
  cell,
  TF,
  Local = FALSE,
  path = "",
  Methylation = FALSE
)
```

Arguments

`NumberofTop`	Number of top co-factors with higher co-binding Percentage to be illustrated
`highestscore`	Co-binding Percentage wich will be the minimum percentage of the shown co-factors.
`cell`	A character string, which is the name of cell under study.
`TF`	A character string which will be the Transcription Factor of interest.
`Local`	A logical value, which will read a local .CSV file in case of TRUE. The file should contain two columns: scores, columnnames which are the co-binding percentages and IDs respectively.
`path`	The path to .CSV file in case Local=TRUE.
`Methylation`	Is a logic argument which indicates if user wants Methylation Score to bu plotted on top of sequence logos or not.

BetaAdder — *A function to merge Beta Score matrices to generate a single matrix.*

Description

This function takes the class object and creates a merge of exclusive Beta Score Matrices by calculating the elementwise weighted average of them; up to the forking position. Weight of each matrix, is the overlapping percentage of intersectPeakmatrix.

Usage

```
BetaAdder(TheObject, sp)
```

Arguments

TheObject	is an object of S4 class that holds original matrices exported from the package TFregulomeR().
sp	is the forking position. User can define up to which position it is required to merge two matrices using this argument.

Value

This function receives a class object, and returns an updated class object.

Examples

```
This function is called within ClassAssignment() function.
```

ConvertToFTRANSFAC — *Generating proper matrix similar to TRANSFAC format of all matrices.*

Description

This function generates a matrix of 5 column (Position,A,T, C, G) with redundant position numbers at Position column reflecting number of leafs and their PWMs.

Usage

```
ConvertToFTRANSFAC(TheObject)
```

Arguments

TheObject	This argument is an object of the class which holds the information ready to be plotted.

Value

This class receives a class Object which holds the plotting data, and updates it by adding the proper matrix of new format: FTRANSFAC.

`Ensembles` *A function for generating the object for TF in different cell lines*

Description

This function browses all the cell lines of a given TF and augoments the provided motif with added cell lines to represent the impact of cell line on motif structure

Usage

```
Ensembles(sp, tfname, tfID, CelllinesNumb)
```

Arguments

`sp`	Forking point for final plot
`tfname`	The name of target tf in strings
`tfID`	the targeted motif using MethMotif IDs as the target cell line under study.
`CelllinesNumb`	Maximum number of cell lines to be considered

`ensemblesfunc` *A function for exporting the motif matrix and augomenting it with additional cell lines*

Description

This function

Usage

```
ensemblesfunc(tfname = "JUN", tfID = "MM1_HSA_K562_JUN")
```

Arguments

`tfname`	The name of target tf in strings
`the`	targeted motif using MethMotif IDs as the target cell line under study.

FPWMPlotter — *A function for generating the forked Position Weight Matrix*

Description

This function takes the generated class object and plots a forked position weight matrix.

Usage

```
FPWMPlotter(TheObject, Methylation = TRUE)
```

Arguments

Methylation	is a logical value. If it set on TRUE, Methylation level chart will also be plotted. If Flase, only sequence logos will be shown.
GraphDataObj	is an object of S4 class with modified and converted data ready to be plotted.

MatrixAdder — *A function to merge motif matrices to generate one matrix as parent node.*

Description

This function takes the object and creates a merge of all matrices by claculating the elementwise addition of them, up to a user specified position (Forking Position).

Usage

```
MatrixAdder(TheObject, sp)
```

Arguments

TheObject	is a object of S4 class that holds original matrices exported from the package TFregulomeR().
sp	is the forking position. User can define up to which position it is required to merge matrices using this argument.

Value

This function recieves a class object, and returns an updated class object by adding merged matrix to parentmatrix slot

Examples

```
This function is called whithin ClassAssignment() function.
```

`ModifyBetaFormat` *A function for converting Beta Score matrices into proper data frames.*

Description

This function receives the S4 class object and converts Betalevel matrices into data frames for better plotting and browsing purposes.

Usage

```
ModifyBetaFormat(TheObject)
```

Arguments

`TheObject`	is an object of S4 class that holds original matrices exported from the package TFregulomeR().

Value

This function receives a class object, and returns an updated class object by modifying Beta Score Matrices.

Examples

```
This function is called within ClassAssignment() function.
```

`ObjectGenerator` *A function togenerate a class object then assign proper data exported from TFregulomeR to its slots.*

Description

This function assigns proper data to their associated slots of a S4 classe. This information is either provided by user, or exported from TFregulomeR's dataware using user specified data.

Usage

```
ObjectGenerator(sp, peak_id_y_list, peak_id_x, height = 2, width = 3)
```

Arguments

`sp`	This argument, defines from which point on, the matrix needs to be forked, or in the other words, up to which point two exclusive matrices need to be aggregated.
`peak_id_y_list`	This argument is a list of TF ID's which will be intersected with Peak_id_x.
`peak_id_x`	This argument holds an id of TFBS compatible with TFregulomeR(). This is the target peak ID wich will be employed by IntersecPeakMatrix of TFregulomeR to extract desired data.
`height`	An argument which allows user to customize the height of final graph relative to screen.
`width`	An argument which allows user to customize the width of final graph relative to screen.

Value

This component, returns a class object which holds all the neccessary information for other functuins.

`PlotMultiFTRANSFACFile` *A function for storing .PDF of plots, by providing a .txt file of FPWMs concatination, in proper format.*

Description

This function reads an stored .txt file of multiple FTRANSFAC matrices and generates the associated plot for each set then stores the figure as a PDF file. Name of each files indicates from each line the infomration is being imported to result to given plot.

Usage

```
PlotMultiFTRANSFACFile(File = "All.txt")
```

Arguments

`File`	the directory of .txt file of multiple FPWMs merged in proper format.

Value

Stores number of PDF files regarding the number of FPWMs provided within the file.

`ReadFTRANSFACFile` *A function for generating a class object from a local file in proper format*

Description

This function reads an stored .txt file of FTRANSFAC format and constructs a class object from it. As default, the returned class Object does not contain Methylation Score matrices. If needed, files exported from TFregulomeR() with the same name and format should be provided before setting MEthylation==TRUE.

Usage

```
ReadFTRANSFACFile(
  File = "MM1_HSA_K562_CEBPB___4-FTRANSFAC.txt",
  Methylation = FALSE
)
```

Arguments

`File`	the directory of .txt file
`Methylation`	a logical argument which indicates if Methylation Score files are provided and needed to be included in Object or not.

Value

A class object for plotting. The Methylation Score matrices can be optionally ommitted or not.

StoreFTRANSFACFile *Generation and storing a file of the standard TRANSFAC format*

Description

This function generates a .txt file of the format TRANSFAC with slight modifications in positions column.

Usage

```
StoreFTRANSFACFile(TheObject)
```

Arguments

TheObject	This argument is an object of the class which holds the information ready to be plotted.IDs, Scores and Froked_PWM are mednatory.

Value

This function stores a .txt file at working directory, and returns name of the file for more convenience.

StoreMultiTRANSFACFile

Generating and storing a .txt file named "All.txt" which contains multiple FPWMs, contatinated together, respecting TRANSFAC format.

Description

This function generates a .txt file which holds number of the data structures needed for one set of plotting in FTRANSFAC format.

Usage

```
StoreMultiTRANSFACFile(List_sp, Listof_peak_id_y_list, List_peak_id_x)
```

Arguments

List_sp	List of forking position numbers for each one of FPWM.
Listof_peak_id_y_list	A list of lists. Each list within this list, is a set of IDs which are going to form one FPWM plot.
List_peak_id_x	A list of IDs. The ID in List_peak_id_x[i] will be employed to form multiple IntersecPeakMatrices with all the IDs exisiting in Listof_peak_id_y_list[i].

Value

This function stores a .txt file at working directory, and returns name of the file for more convenience.

`storeTRANSFAC` *A function for storing TRANSFAC files of the forked matrices.*

Description

This function generates files of regular TRANSFAC format in order to further analysis and evaluation. Each file name holdes the name of Transfactor of interest, and the co-factor that is under analysis in the current matrix.

Usage

```
storeTRANSFAC(TheObject)
```

Arguments

`TheObject`	the input is the object of FPWM class that holds the raw matrices directly exported from TFregulomeR().

`ToTFBSTools` *A function for generating object of TFBStools holding PFM of each fokred matrix.*

Description

This function generates a TFBStools object of each matrix present in FPWM class object, and returns a list containing all the objects.

Usage

```
ToTFBSTools(TheObject)
```

Arguments

`TheObject`	is the object of the FPWM class. It needs to contain the matrices, IDs and parent matrix.